Psychological Review
2014, Vol. 121, No. 3, 337–366

© 2014 American Psychological Association
0033-295X/14/$12.00 http://dx.doi.org/10.1037/a0037015

Opponent Actor Learning (OpAL): Modeling Interactive Effects of Striatal Dopamine on Reinforcement Learning and Choice Incentive

Anne G. E. Collins and Michael J. Frank
Brown University

The striatal dopaminergic system has been implicated in reinforcement learning (RL), motor performance, and incentive motivation. Various computational models have been proposed to account for each of these effects individually, but a formal analysis of their interactions is lacking. Here we present a novel algorithmic model expanding the classical actor-critic architecture to include fundamental interactive properties of neural circuit models, incorporating both incentive and learning effects into a single theoretical framework. The standard actor is replaced by a dual opponent actor system representing distinct striatal populations, which come to differentially specialize in discriminating positive and negative action values. Dopamine modulates the degree to which each actor component contributes to both learning and choice discriminations. In contrast to standard frameworks, this model simultaneously captures documented effects of dopamine on both learning and choice incentive—and their interactions—across a variety of studies, including probabilistic RL, effort-based choice, and motor skill learning.

Keywords: dopamine, striatum, reinforcement learning, choice incentive, computational model

Dopamine plays a crucial role in human and animal cognition, substantially influencing a diversity of processes including reinforcement learning, motivation, incentive, working memory, and effort. Dopaminergic neurons in the substantial nigra and ventral tegmental area project to a very wide set of subcortical and cortical areas, with strongest innervation in the basal ganglia (BG), specifically in the ventral and dorsal striatum. Dysregulation of dopamine is present in a wide array of mental illnesses such as Parkinson's disease, attention-deficit/hyperactivity disorder (ADHD), schizophrenia, and Tourette's syndrome and is a central pharmaceutical target used to treat symptoms across these and numerous other pathologies.

Although considerable progress has been made in our understanding of its various distinct roles, there remain fundamental debates concerning its precise mechanisms and functions, especially regarding their integration and interactions. In reward-based decision making in particular, two largely separate traditions have studied the reinforcement learning and the incentive theories of dopamine (Berridge, 2007). Despite solid evidence for both theories, theoretical and empirical studies tend to favor and focus on one or the other interpretation, with little attempt to unify them or to study their interaction. Here we develop an explicit computational analysis of the dual role of striatal dopamine in modulation of incentive motivation (affecting choice), reinforcement learning, and how these processes interact. This endeavor allows us not only to account for both types of findings alone but also those that could not be explained by either theory in isolation.

RL Theory of Dopamine

One widely accepted theory of dopamine function relates to its role in model-free reinforcement learning (RL). Specifically, phasic firing of midbrain dopamine neurons convey reward prediction errors that facilitate plasticity in the striatum (Montague, Dayan, & Sejnowski, 1996; Schultz, 1997). Many studies have since provided strong support for this notion (Arias-Carrión, Stamelou, Murillo- Rodríguez, Menéndez-González, & Pöppel, 2010; Bayer & Glimcher, 2005; Bayer, Lau, & Glimcher, 2007; Nakahara, Itoh, Kawagoe, Takikawa, & Nikosaka, 2004; Nomoto, Schultz, Watanabe, & Sakagami, 2010). Reinforcement learning models have been routinely used to account for dopaminergic modulation of behavioral and neural signals during learning tasks (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Jocham, Klein, & Ullsperger, 2011; McClure, Daw, & Read Montague, 2003; O'Doherty et al., 2004; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; Samejima, Ueda, Doya, & Kimura, 2005; Schönberg, Daw, Joel, & O'Doherty, 2007).

Such models assume that each action has a single value, which gets incremented or decremented by dopamine-encoded reward prediction errors to drive learning. Choice between different actions is accomplished by comparing the current action values among all the available actions in the given sensory state, and stochastically choosing one, such that actions with higher values

Anne G. E. Collins and Michael J. Frank, Department of Cognitive, Linguistic and Psychological Sciences, Brown Institute for Brain Science, Brown University.

This project was supported by National Science Foundation Grant 1125788 and National Institute of Mental Health Grant R01MH080066-01. The views stated in this article do not necessarily represent those of the funding institutions. Code supporting simulations found in this article can be downloaded from http://ski.clps.brown.edu/CollinsFrankOpAL_code.zip

Correspondence concerning this article should be addressed to Anne G. E. Collins or Michael J. Frank, CLPS, Brown University, 190 Thayer Street, Providence, RI 02912. E-mail: Anne_Collins@Brown.edu or Michael_Frank@Brown.edu

are more likely to be selected. These models account for a wide variety of data but alone cannot capture the apparent modulatory influence of dopamine on incentive choice—the tendency to differentially weigh costs and benefits—after learning has occurred (Berridge, 2012; Salamone, Correa, Mingote, & Weber, 2005). They also don't easily accommodate the asymmetrical influences of dopamine manipulations on learning from positive versus negative outcomes. Conversely, theories and models of incentive choice (Zhang, Berridge, Tindell, Smith, & Aldridge, 2009) do not account for progressive learning reinforcement effects or the findings that motor symptoms of Parkinson's disease can progress even without further dopaminergic degeneration (Beeler, Frank, McDaid, & Alexander, 2012).

In contrast, neurobiology and neural network models thereof suggest a more complex dual opponent system for action and learning. Dopamine (DA) is known to modulate activity and plasticity of striatal medium spiny neurons (MSN) in two separate populations of cells that project to different BG output nuclei (Frank, 2005; Gerfen, 2000; Shen, Flajolet, Greengard, & Surmeier, 2008; Surmeier, Ding, Day, Wang, & Shen, 2007). Striatal MSNs originating in the direct (striatonigral) pathway predominantly express dopamine D1 receptors and act to facilitate actions (Kravitz et al., 2010). By stimulating D1 receptors in these neurons, dopamine enhances the signal-to-noise ratio and amplifies activity and plasticity (long term potentiation). By contrast, striatal MSNs originating in the indirect (striatopallidal) pathway predominantly express dopamine D2 receptors, and act to suppress actions. By stimulating D2 receptors in these neurons, dopamine inhibits their activity and induces long term depression. Thus overall, increases in dopamine act to preferentially emphasize processing in the D1 facilitatory pathway and to suppress processing in the D2 suppressive pathway, whereas decrease in dopamine have the opposite effect, potentiating the D2 pathway. This has been proposed as the mechanism by which DA promotes approach learning in the direct pathway and avoidance learning in the indirect pathway (Frank, 2005), with opposite-coding but apparently redundant representations of action values and learning. Although the direct (D1-MSNs) and indirect (D2-MSNs) pathways are often labeled as *Go* and *NoGo* pathways due to their link to approach and avoidance, respectively, in the models they do not just encode a message signaling to go or not, but rather the aggregated evidence in favor of each action versus against that action. The final choice is a function of the relative differences in the amounts of evidence for each action considered, via competition at all stages in the corticostriatal circuit.

Dopamine dysregulation thus acts in opposite directions in the separate pathways. This feature has been widely used to exhibit effects of dopamine-related drugs, genes, pathologies, etc., all of which act to induce an asymmetry in the treatment of positive versus negative outcomes (e.g., by having opposite effects on approach vs. avoidance learning). As an example, nonmedicated Parkinson's patients have naturally low dopamine levels and exhibit better learning from negative than positive reward prediction errors, whereas the same patients while taking dopaminergic medication show better learning and choice based on positive outcomes but worse performance in avoiding negative outcomes (Bódi et al., 2009; Cools et al., 2009; Frank, Moustafa, et al., 2007; Frank, Seeberger, & O'Reilly, 2004; Moustafa, Sherman, & Frank, 2008; Palminteri, Boraud, Lafargue, Dubois, & Pessiglione, 2009; Smittenaar et al., 2012). Similar effects of dopamine manipulations have been observed in healthy and other populations (Cools et al., 2009; Frank, Moustafa, et al., 2007; Frank, Santamaria, Reilly, & Willcutt, 2007; Jocham et al., 2011; Pessiglione et al., 2006). This reinforcement learning theory of dopamine function can also account for other counterintuitive phenomena, such as aberrant learning in some situations (e.g., Beeler, Daw, Frazier, & Zhuang, 2010; Wiecki, Riedinger, von Ameln-Mayerhofer, Schmidt, & Frank, 2009: learned catalepsy), and provides a mechanism explaining progression of Parkinson's disease symptoms even without further dopaminergic degeneration (Beeler et al., 2012).

More direct probing of the role of dopamine in specific neural circuits comes from optogenetic studies confirming a role for the D1 and D2 pathways in approach and avoidance learning (Kravitz, Tye, & Kreitzer, 2012). After a specific action was selected endogenously by a mouse, optogenetic stimulation of D1 MSNs resulted in positive reinforcement of that specific action, causing the mouse to repeat it in the future. Conversely, optogenetic stimulation of D2 MSNs caused that action to be avoided. Notably, the effect of stimulation was applied only following the choices in these studies, such that any subsequent change in behavioral preferences can only be attributed to a learning mechanism, rather than a direct performance effect. Moreover, the effect of stimulating D1 and D2 cells mimics that which would occur as a result of dopamine bursts and dips, respectively. While this study shows that D1 and D2 stimulation is sufficient to induce approach and avoidance learning, respectively, other genetic engineering studies also show that they are necessary (Hikida, Kimura, Wada, Funabiki, & Nakanishi, 2010). Thus, many independent data-points confirm the role of dopamine on the striatum in reinforcement learning, either indirectly or directly. However, many reinforcement learning studies also fail to control for potentially confounding incentive effects of dopamine, as described next.

Incentive Theory of Dopamine

Outside of the field of reinforcement learning, various types of evidence indicate that dopamine is also involved directly in choice, with links to motivation, incentive, vigor or effort willingness (Berridge, 2012; Smith, Berridge, & Aldridge, 2011; Wassum, Ostlund, Balleine, & Maidment, 2011). Numerous studies have shown, for example, that hyperdopaminergic rats were willing to work more for identical amount of reward (Beeler et al., 2010; Cousins & Salamone, 1994; Salamone et al., 2005). The effective apparent "cost" of effort is bidirectionally modulated by manipulation of indirect D2 MSN activity: Pharmacological manipulations that enhance such activity result in more avoidance of effortful actions, whereas inhibition of this pathway has the opposite effect, decreasing the effective cost (Farrar et al., 2010, 2008; Mingote et al., 2008; Nunes et al., 2010). Neural models suggest that these effects are mediated by differential coding of positive and negative consequences of actions in distinct MSN populations, as observed in electrophysiological studies (Samejima et al., 2005).

Recent optogenetic work (Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012) has also confirmed more precisely that specific action values can be inflated or diminished by stimulating D1 or D2 MSNs during the choice period (as opposed to during the outcome in the learning study described previously). Stimulating

D1 MSNs in one hemisphere acted to increase the likelihood of choosing the contralateral action. Notably, this was not a pure motor effect: Stimulation did not simply deterministically increase motor responding but, rather, acted to boost action value. Higher levels of stimulation were needed to induce choice for actions that had low learned values and lower levels for actions already having high values. Strikingly, D2 MSN stimulation had the opposite effect, effectively decreasing the action value (or increasing its effective cost). In summary, when applied at the time of choice, stimulation of D1 (respectively, D2) mimicked an additive positive (negative) effect on the action's recent estimated value.

Other modeling studies have proposed that tonic dopamine modulates response vigor to optimize reward (or avoid punishment) per unit time (Dayan, 2012; Niv, Daw, Joel, & Dayan, 2007). However, note that these models have only considered effects on vigor (i.e., the speed of response execution or how hard to work) but not the effects on choices between actions with difference valences/incentives. Moreover, they focus on effects of increased DA signaling and not the relative enhancement of performance in some cases with DA depletion.

Recent studies have debated the link between incentive or performance effects of dopamine on one hand, and reinforcement learning effects on the other, with some arguing that all reinforcement learning effects could be reinterpreted in terms of incentive salience (Berridge, 2012). Indeed, many of the above-described experiments demonstrating differential influence of dopamine manipulations to positive versus negative outcomes have not dissociated between learning versus incentive accounts. Arguably, some of the asymmetry in reward versus punishment learning in Parkinson's disease and other human studies could potentially be accounted for by differential incentive at the time of choice, even given symmetrical learning. For example, some recent work provides evidence that dopamine modulations can influence relative sensitivity to positive versus negative outcomes in action selection, when learning effects were not observed or were not possible in the task (Shiner et al., 2012; Smittenaar et al., 2012). Specifically, Smittenaar et al. (2012) showed that Parkinson's patients on compared to off medication were better able to select actions that would lead to the most rewarding outcomes, even when there was no differential values assigned to stimulus outcomes during learning itself, where reward values were only assigned to the outcomes after learning had taken place. Conversely, Shiner et al. (2012) used a standard stimulus-value learning procedure, but then manipulated DA medications only after learning, and nevertheless observed that patients on medication during this postlearning phase exhibited better performance on rewarding than aversive choices. These studies hint at the presence of a performance/incentive effect but do not exclude an additional role of dopamine in learning. Indeed, the incentive effect in these studies cannot explain all the previously documented data: perhaps the most robust effect of dopamine elevations across human studies is to impair learning from negative outcomes, whereas in these studies there was only an effect on sensitivity to positive outcomes, and the magnitude of the overall medication effect on the differential sensitivity to positive versus negative outcomes was substantially more modest than that observed in the various studies that could also have been influenced by learning. Moreover, other studies provide evidence for learning effects, for example, striatal response to reward prediction errors during learning are predictive of subsequent reward-based choice preferences, and this relationship is modulated by dopaminergic manipulation (Jocham et al., 2011). In sum, current evidence implies that dopaminergic manipulation influences both learning and incentive.

While there remains some debate over the respective contributions of learning and performance effects of dopamine on various behavioral measures, some studies provide evidence for interactions between these two functions. In particular, Beeler et al. (2012) used dopamine antagonists that induced performance deficits in rodents confronted with a motor skill task. By themselves, these effects converge with a long history of evidence that striatal dopamine is important for motor performance, as in Parkinson's disease. Notably, however, this study showed that even after drug washout, animals were slower to acquire the correct motor skills compared to naive animals who had never been exposed to the task, and compared to animals who had also been administered dopamine antagonists but not paired with the task. This study demonstrated that the drug effects on performance induced an "aberrant learning" process causing animals to learn to avoid selecting the actions that would have been adaptive. Subsequent experiments demonstrated similar effects when D2 blockade was applied after learning of an established skill: In this case, performance did not degrade immediately but, rather, progressively declined, consistent with an induction of aberrant learning, and with parallel synaptic plasticity studies showing that D2 antagonism enhanced potentiation of striatopallidal synapses (Beeler et al., 2012). These effects are also coherent with other evidence that moderate doses of D2 antagonists can induce progressive Parkinsonian symptoms in the form of catalepsy sensitization (Amtage & Schmidt, 2003; Klein & Schmidt, 2003), and both of these effects are accounted for by simulations of D2 antagonism in neural models (Beeler et al., 2012; Wiecki et al., 2009) These studies highlight a role for interactions between performance effect of dopamine (in this case, a lack of dopamine), which induce learning effects that then further exaggerate performance effects, etc.

Here we present a new reinforcement learning model that allows us to simultaneously account for incentive, learning and interaction effects of dopamine. We aim to provide a theoretically simple algorithmic model, with parameters and variables that can easily be related to biologically interpretable measures of interest, such as tonic or phasic dopamine level, D1 and D2-expressing striatal neuronal activity or synaptic strengths, synaptic plasticity etc. Our approach is inspired by two distinct levels of modeling: on one hand, the well-known and widely used actor-critic algorithm (Sutton & Barto, 1998); on the other hand, the more biologically detailed neural network description of corticobasal ganglia loops including multiple pathways (Frank, 2005). Although previous attempts to link these levels of modeling exist, these did not consider separate valuation systems for action selection and learning, or the effects of incentive but, rather, included a single value for each action and only allowed for asymmetric learning rates for positive versus negative prediction errors (Doll, Hutchison, & Frank, 2011; Frank, Moustafa, et al., 2007). As shown below, this mechanism is insufficient to account for the range of data. The aim here is to provide a model that can exhibit more generally distinct but interacting motivational incentive, performance and learning effects, which can account for a range of findings in the literature that are not accommodated by existing formulations. We further provide an analysis proposing a normative reason for this separation of valuation systems.

Model and Simulation Methods

OpAL Model Description

Our new model, labeled OpAL for Opponent Actor Learning, relies on an actor-critic architecture. The actor-critic architecture assumes that one system, the critic, estimates the values of the current state of the environment, whereas the actor selects actions. When outcomes are better or worse than expected, the critic generates a reward prediction error, which is used for two purposes: to update its future estimates, so that it is a better estimate of the value, and to modify the actor weights. Actions that produce positive prediction errors are reinforced, whereas those that produce negative prediction errors are punished. Commonly, the critic is assigned to ventral striatum and the amygdala (Hazy, Frank, & O'Reilly, 2010; O'Doherty et al., 2004), whereas the actor is considered to be instantiated by dorsal striatal interactions with pre/motor cortex.

Critic learning. The critic in our model is similar to that in classical formulations (but see Discussion). It estimates the expected value of a given choice option[1] and updates this value through a simple delta rule learning algorithm:

$$V(t+1) = V(t) + \alpha_C \times \delta(t). \tag{1}$$

Thus, the update of the estimated value V is proportional to the prediction error $\delta(t) = r(t) - V(t)$, where $r(t)$ indicates reinforcement received at time t, and α_C is the critic learning rate. We make the common assumption that critic values are represented in ventral striatum and that phasic signals of dopamine convey the critic prediction error (Dayan & Daw, 2008; Montague et al., 1996; Roesch, Calu, & Schoenbaum, 2007).

Actor learning. The typical actor selection mechanism assigns a set of weights to each action, and increments or decrements these weights as a function of critic prediction errors. In the OpAL model, we separate the actor into two sets of weights, representing corticostriatal synaptic weights into direct (G for Go) and indirect (N for NoGo) MSN populations coding for state-action pairs (s, a). For simplicity of exposition, we consider here a single state with multiple action choices, thus simplifying (s, a) to a. These actor weights, labeled, respectively, $G_a(t)$ and $N_a(t)$, are constrained to be positive (firing rates and glutamatergic synaptic weights cannot be negative).

Learning for these actor weights mimics learning mechanisms in neural network-models as follows:

$$G_a(t+1) = G_a(t) + [\alpha_G G_a(t)] \times \delta(t) \tag{2}$$

$$N_a(t+1) = N_a(t) + [\alpha_N N_a(t)] \times [-\delta(t)] \tag{3}$$

Here, $\delta(t)$ is the previously defined critic prediction error and α_G and α_N are learning rates for Go and NoGo weights, respectively.

This model structure (dual actor weights) and its temporal dynamics (update rules) reflect a departure from typical RL models. First, the presence of separate G and N weights as well as the two unusual features of the update rules are biologically motivated. The first feature of the update rule is that contrary to G-weights, N-weights are updated through the opposite sign of the prediction error.[2] This captures the notion that dopamine has opposite effects on plasticity via stimulation of D1 and D2 receptors in the different populations but that they both can undergo potentiation and depression. Intuitively, G-weights accumulate prediction errors, so should come to represent an index of how good an option is, while N-weights increase with negative prediction errors and decrease with positive ones, so should come to represent how aversive an option is.

The second feature of the update rule that departs from typical RL updating, is that the extent of learning is determined not only by the prediction error but also by the current actor weight, as a multiplicative factor on the learning rate. This captures the often quoted *three-factor Hebbian rule*, where learning depends on presynaptic activation (from cortex, the stimulus and action), postsynaptic activation in striatum (proportional to the actor weight), and dopamine (Reynolds, Hyland, & Wickens, 2001). Although this rule is often linked to reinforcement learning models, those typically do not actually implement a three-factor rule. Indeed they effectively incorporate presynaptic (stimulus-action representations) and dopamine (prediction error) constraints in learning rules but are not contingent on postsynaptic modulation, which we obtain here by incorporating the actor weight (which would determine the level of postsynaptic activation). Importantly, simulations below compare this update rule to one in which the actor weight is absent from the update equation and show that the modulation of learning by actor values is necessary to account for the range of data, including the effects of performance on learning, and the tendency for G and N weights to preferentially discriminate between positive and negative action values, respectively.

Another departure from typical RL models, the presence of G and N weights, rather than actor weights within a unitary system, potentially affords computational advantages in terms of flexibility, by allowing separate dynamical regulation of whether G or N should affect choice (see next section). Last, we show in the results section that these features are critical to account for data that show that dopamine can affect incentive without affecting learning, and vice versa.

Policy. Choice between different options, for example between different available action choices a_i, is given as a softmax choice policy on the linear combination of the actor weights:

$$Act_a(t) = \beta_G G_a(t) - \beta_N N_a(t) \tag{4}$$

$$p(a) = \frac{e^{Act_a(t)}}{\sum_i e^{Act_{a_i}(t)}} \tag{5}$$

Here, $p(a)$ is the probability of choosing action a. It depends on the combined actor weight Act_a, that is the weighted difference

[1] This could be a single state s, or a state-action pair (s, a), following indications that prediction error could correspond to state-action prediction errors (Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006; Roesch, Calu, & Schoenbaum, 2007).

[2] Note that this is $+\delta$ versus $-\delta$, so for example N weights increase with negative prediction errors and decrease with positive prediction errors. This implementation contrasts with previous attempts to model asymmetric learning via separate learning rates to rectified prediction errors, i.e., that only apply when δ is positive or negative, respectively, e.g., in Frank, Moustafa, Haughey, Curran, and Hutchison (2007).

between G_a and N_a, compared to that for all other candidate actions, representing competition between direct and indirect pathway activity in the output nucleus of the basal ganglia, GPi (see Figure 1 left). Parameters β_G and β_N modulate the extent to which the G and N weights are represented in a given trial, such that $\beta_G G_a(t)$ represents the activation of the associated G population and $\beta_N N_a(t)$ represents the activation of the associated N population. The softmax function implements the nonlinearity in choice as a function of value. Thus, depending on the asymmetry in β_G versus β_N, proposed below to relate to dopamine levels at time of choice, the benefits and costs of actions are differentially represented, and different actions can be selected (Figure 1). Note that we can rewrite the parameters as $\beta_G = \beta \times (1 + \rho)$ and $\beta_N = \beta \times (1 - \rho)$. In this form, $-1 < \rho < 1$ represents the asymmetry between the weights and β corresponds to the classic softmax inverse temperature parameter, controlling exploration versus exploitation.

Reaction time. We model the reaction time of a choice a as a function of its actor value Act_a through softmax:

$$RT(a) \propto RT_0 + 1/(1 + e^{(Act_a(t)-\theta)}), \qquad (6)$$

where θ is a threshold for Go relative to NoGo pathway activity needed to facilitate the action, and RT_0 is simply baseline reaction time. This equation captures the RTs generated by neural network simulations in which relatively greater Go than NoGo population activity results in faster RTs (Moustafa et al., 2008; Wiecki et al., 2009) and where the decision threshold (as estimated through a drift diffusion model) is in part controlled by the output of the basal ganglia, corresponding to the Act value here (Ratcliff & Frank, 2011).

Simulating Dopaminergic Effects on Learning and Incentive

Simulations reported below show that the model as defined above can capture potential asymmetries in both learning and choice incentive. Regarding learning, in neural circuit models dopaminergic modulations can enhance phasic burst signaling, enhancing sensitivity to positive prediction errors. However, higher tonic levels can also prevent D2 receptors from detecting phasic dips, thereby reducing sensitivity to negative prediction errors (Frank, 2005). Conversely, low dopamine levels may reduce phasic burst signaling but actually increase the sensitivity to dips (Frank & O'Reilly, 2006). In the OpAL framework, this modulation can be modeled by asymmetries in the actor learning rate parameters α_G and α_N, capturing the sensitivity of D1 and D2 MSNs to dopaminergic signals and resultant effects on plasticity. Conversely, choice incentive effects of dopamine manipulations at the time of choice can be modeled by asymmetries in β_G and β_N, modulating the degree to which learned values in the two pathways are expressed. For example, a high level of dopamine at the time of choice enhances active D1-MSNs, but inhibits D2-MSNs, so would be modeled by an increase in β_G and decrease in β_N, and conversely for a decrease in dopamine.

To summarize, as a first approximation, we model potential *learning* effects with the α_G, α_N parameters, and potential *incentive* or *performance* effects with the β_G, β_N parameters. Broadly, this distinction accords with the separate effects of phasic dopamine encoding prediction errors (Montague et al., 1996), compared with the tonic (baseline) effects of dopamine relating to vigor (Niv et al., 2007). However, we note that our model suggests that the key distinction between incentive and learning effects simply depends on the level of striatal dopamine at the time of choice versus the time of reinforce-

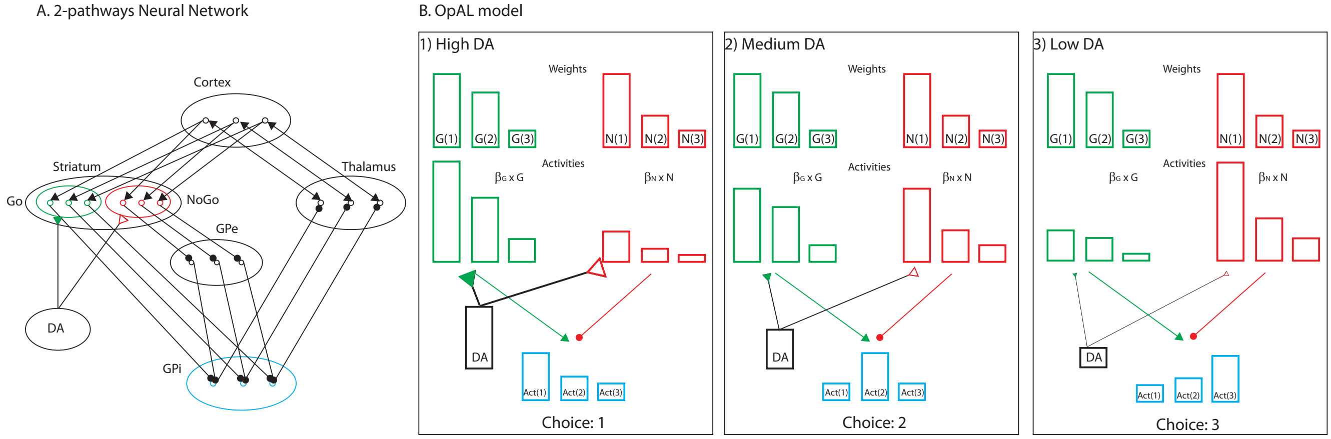

Figure 1. Neural network and opponent actor learning (OpAL) models. A. Schematic depiction of the neural network representing corticobasal ganglia loops, used in Frank (2005) and others to simulate various effects of dopamine on learning and performance. B. Representation of OpAL model, with example of choice between three options for a given stimulus, with three different dopaminergic states. In a normal dopaminergic state (2, middle), the model's actor weights (Act) favor an action that has high G weights and low N weights, relatively to the other options. In a high dopaminergic state (1, left), G values are emphasized more than N values on actor choice are emphasized, leading to the choice of an action that has relatively highest G weights, with little regard for the action costs. In contrast, in a low dopamine state (3, right), the reverse happens, and the model chooses action that has lowest N weights. DA = dopamine; GPI = internal segment of the globus pallidus. See the online article for the color version of this figure.

ment, respectively, and hence any phasic bursts that occur during choice would also affect incentive choice and reaction times (see, e.g., Satoh, Nakai, Sato, & Kimura, 2003, in which phasic DA signals were related to faster RTs), captured by β_G and β_N parameters.

Results

Model Dynamics

Standard reinforcement learning models, including actor-critic models, have been proven, under reasonable assumptions, to converge in probability to estimating the expected sum of future discounted reward given the state and/or action. *It is desirable to ensure that Act in our model has similar properties that are useful for rational learning and decision making*, for example, that it does not diverge or that it is a monotonically increasing function of expected reward. We show here some simulations that validate the most critical aspects to ensure that OpAL defines a reasonable learning and choice policy (and include some theoretical derivations in supporting information).

In a first set of simulations (see Figure 2), we manipulated expected value of a choice by parametrically changing reward value $r > 0$ and probability of reward versus no reward ($r = 0$), in a single forced choice setting, using neutral (symmetrical) parameters ($\alpha_G = \alpha_N = \alpha_C = 0.1$; $\beta_G = \beta_N = 1$). These simulations show that *G* and *N* are, respectively, increasing and decreasing convex functions of *r* and $p(r)$, while *Act* is a nonlinear increasing function (Figure 2B).

In particular, *G* weights are positively correlated with true expected value, which increase approach tendencies, but the convexity of the curve indicates that the function is nonlinear: they exhibit greater differentiation among higher value stimuli. Thus, a fixed difference of ε between the probability of reward of two stimuli/actions (i.e., $p(r)$ and $p(r + \varepsilon)$) will be amplified in *G* weights to a greater extent when $p(r)$ is high compared to low, especially if *r* is also high.

In contrast, *N* weights are negatively correlated with true expected value and act to support avoidance tendencies. Here, the convexity indicates that *N* weights differentially emphasize lower (rather than higher) value stimuli/action representations. These effects are particularly visible in the time course plots (Figure 2A, middle graphs for *G* and *N*). However, it should be noted that with symmetrical parameters as used here, the net *Act* values evolve without biases present in *G* and *N*, like standard reinforcement learning values but with emphasized differentiation at extreme values (compare top and bottom graphs).[3]

Parameter Effects

Actor learning rate (α_G and α_N) effects. In a second set of simulations (see Figure 3A), we separately manipulated the actor learning rates. These findings indicated that greater learning rates emphasize the modulation seen in normal dynamics, such that with increasing α_G, *good* options are perceived with even stronger *G* weights, and *bad* options with even smaller *G* weights, leading to an exaggerated influence of reward value in actor weights. Conversely, increasing α_N results in greater representation and differentiation among low valued options. Note that these effects hold despite the fact that positive and negative prediction errors are treated identically in the update of both *G* and *N* weights, i.e., that positive prediction errors increase *G* and decrease *N* weights, and vice versa for negative prediction errors. The reason these different systems differentiate among positive versus negative outcomes lies in the Hebbian modulation of learning rules, which differentially impact the accumulation of reward prediction errors across time, such that they come to represent positive and negative values.

Choice incentive (β_G and β_N) effects. We saw earlier that *G* weights amplify differences in $p(r)$ and in *r* more as their values increase, whereas *N* weights amplify differences in $p(r)$ and *r* more as their values decrease. Here we show that modulations of β_G versus β_N parameters, simulating dopaminergic manipulations at time of choice, further magnify the corresponding weight biases. When $\beta_G = \beta_N$, the differential emphasis of high or low values cancels out perfectly in *Act* weights. However, shifting the balance between the *G* and *N* systems using asymmetric βs reveals the influence of the corresponding bias—even in the presence of symmetrical learning.

Indeed, simulations (Figure 3C) showed that with $\beta_G > \beta_N$, *Act* is a convex function of $p(r)$, leading to greater differentiation among *good* than *bad* options (revealing influence of *G*-weights), whereas the converse was true with $\beta_N > \beta_G$, revealing the influence of *N*-weights.

Simulating Optogenetic Effects on Learning and Performance

The above simulations reveal how changes in environmental contingencies (reward probabilities and magnitudes) influence model dynamics. These modeling results can be directly linked to capture findings from optogenetic studies showing that stimulation of D1 or D2 MSNs can differentially influence incentive choice or reinforcement learning depending on whether stimulation is delivered during choice or outcome (Kravitz et al., 2012; Tai et al., 2012). We model these experiments explicitly here and reproduce all the main findings.

Kravitz et al. (2012) stimulated striatal MSNs expressing either D1 or D2 receptors immediately following the rat's selection of a particular action. D1 stimulation induced approach learning, such that this action was preferentially selected in the future, whereas D2 stimulation induced avoidance learning. Moreover, these effects of direct MSN stimulation did not depend on dopamine, as they persisted when dopamine blockers were administered.

We model the optogenetic stimulation of MSNs in OpAL by enhancing the activity-dependent learning rule for the corresponding population (i.e., optogenetic simulations were additive to G values for D1 MSNs stimulation and to *N* values for D2 MSNs stimulation). Specifically, we assumed that stimulation influenced learning in the same way that dopaminergic prediction error does, by modulating the *G* and *N* activity levels:

$$G_a(t+1) = G_a(t) + [\alpha_G G_a(t)] \times [\delta(t) + Opt_G] \qquad (7)$$

$$N_a(t+1) = N_a(t) + [\alpha_N N_a(t)] \times [-\delta(t) + Opt_N] \qquad (8)$$

with $Opt_G = Opt$ when D1 MSNs are stimulated (and 0 otherwise),

[3] Note that actor weights are initialized to 1 throughout this article, unless explicitly stated otherwise but that no results are dependent on this. Any positive initialization would produce identical results (see Appendix).

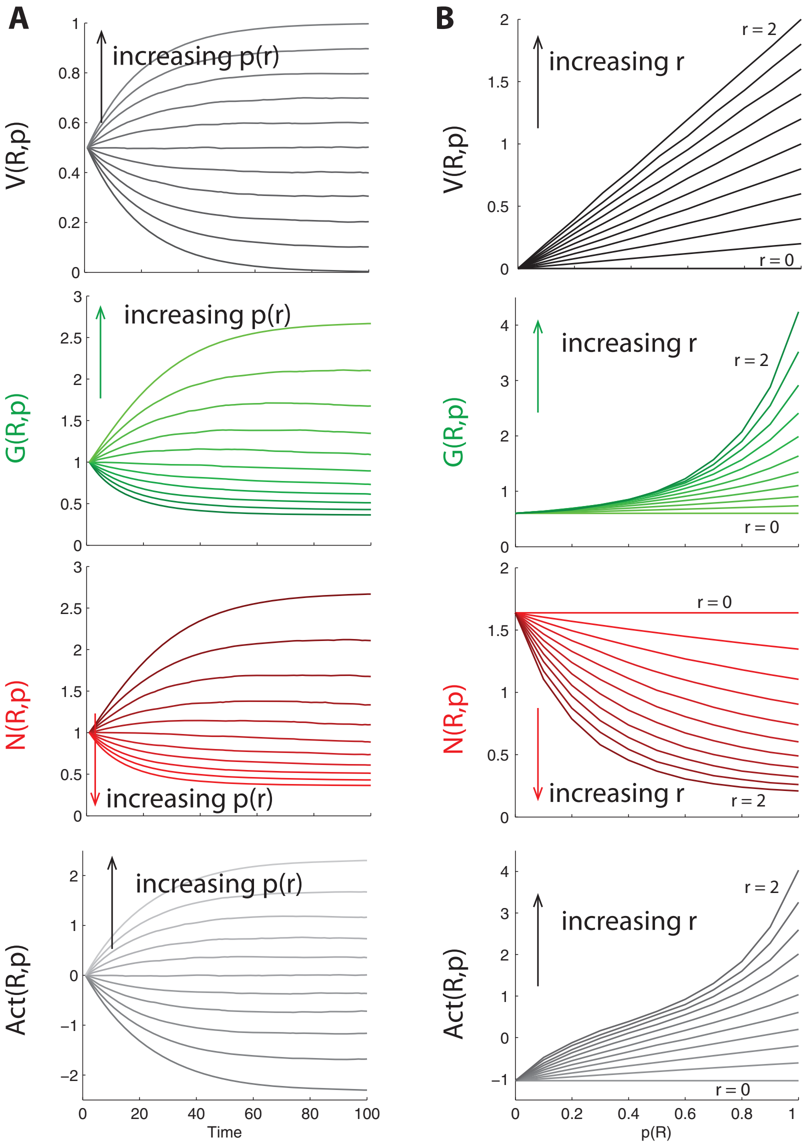

Figure 2. Model dynamics as a function of time, reward value and reward probability. All values are averaged over 1,000 simulations. Model values are (from top to bottom) critic values *V*, *G* weights, *N* weights and actor values $Act = G - N$. A. In these simulations, we show evolution of the model variables as a function of time for different probabilities of reward, with symmetrical model parameters ($\alpha_G = \alpha_N = \alpha_C = 0.1$; $\beta_G = \beta_N = 1$). The top graph shows that critic value *V* quickly converges to true expected value. The second and third graphs show opposing evolution of actor weights *G* and *N*, in opposite directions as a function of reward value R and $p(r)$. Note that *G* weights exaggerate differences in high expected values, whereas *N* weights exaggerate differences in low expected values. The bottom graph shows that with symmetrical β parameters, the actor weights *Act* attain values positively correlated to true expected values, without bias, but that this representation of expected value is nonlinear. B. In these simulations, we show final values after 100 trials, with manipulation of reward value *r* and reward probability $p(r)$. See the online article for the color version of this figure.

and $Opt_N = Opt$ when D2 MSNs are stimulated (and 0 otherwise). $Opt > 0$ is the parameter representing the strength of the stimulation. Since no primary reinforcement was provided in this experimental paradigm, there was no value learning in the critic. Thus, phasic dopamine prediction errors were assumed to be $\delta(t) = 0$, but we allowed for random fluctuations by adding 0-mean Gaussian noise. We also assumed some forgetting in learned weights with rate $\phi = 0.2$, to capture overall extinction effects (but with no asymmetry in the *G* and *N* weights) as well as undirected noise in choice selection ($\varepsilon = 0.25$). Note that none of the results are qualitatively dependent on these parameters. Simulations were run with symmetric learning rate $\alpha_G = \alpha_N = 0.1$, as well as symmetric softmax weights $\beta_G = \beta_N = 10$, and optogenetic strength $Opt = 0.2$, for 100 iterations. To obtain similar

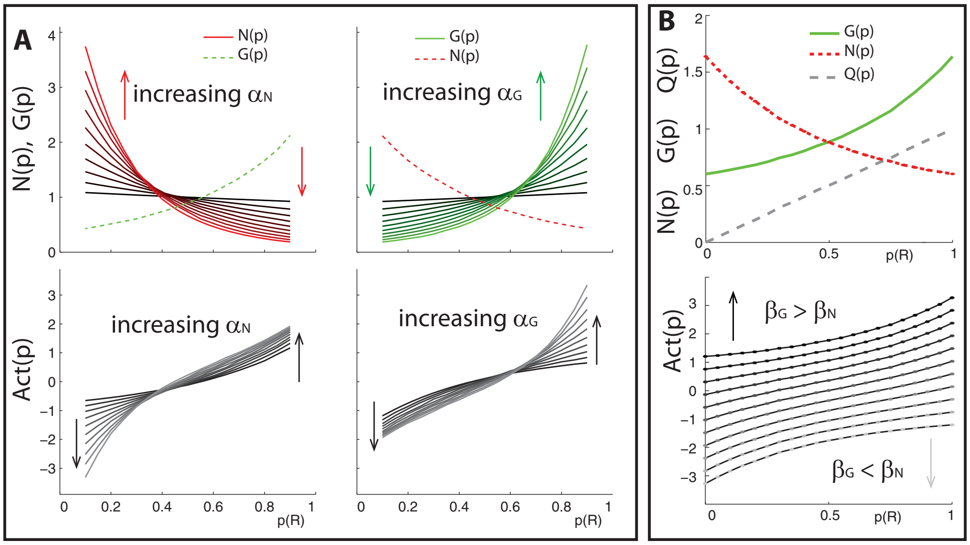

Figure 3. Model values as a function of reward probability p(R) and model parameters. All values are final values after 100 trials, averaged over 1,000 simulations. Model values are (from top to bottom) critic values *V*, *G* weights, *N* weights and actor values *Act* = *G* – *N*. A. Learning rates α_G and α_N (increases in parameter values are represented by lighter lines and arrow direction). Increase in α_G (right) enhances the coded values of, and discrimination among, *good* options and depresses those of bad ones, and vice versa for α_N (left). B. β_G and β_N parameters. With $\beta_G = \beta_N$, the actor weight *Act* is a linear function of expected value. However, inducing a weight asymmetry toward β_G (darker dots) induces an increase in *Act* value (more willingness to choose) and convexity in its representation (better differentiation of *good* options). The opposite happens for $\beta_N < \beta_G$ (lighter circles). See the online article for the color version of this figure.

number of trials to those observed in the experiment, we modeled reaction times with RT = 5 + 10/(1 + *exp*(*Act*)) sec. Simulations were run to provide 30 min sessions, corresponding to the experimental design in Kravitz et al. (2012). Effects of dopamine blockade were simulated by setting an asymmetry in choice parameters, with $\rho = -0.3$, such that $\beta_G = \beta * (1 + \rho)$ and $\beta_N = \beta * (1 + \rho)$.

Accordingly, in the first set of simulations including only stimulation to the D1 or D2 MSNs, the model developed stronger G weights for the action that was associated with the stimulation side given D1 MSN stimulation and thus increased its likelihood of selection (blue bars in Figure 4, middle); this was true of the first few trials of each session, indicating a learning, not performance effect (Figure 4, left). For D2 MSNs stimulation, the model developed stronger *N* weights associated with the triggered action, thus learned avoidance. Furthermore, these increased *N* weights were accompanied by slower reaction times, thereby leading to fewer overall choices made within the same period of time, much like that observed in rodents (red bars in Figure 4, right). Notably, these findings persisted even in the presence of simulated DA blockade (simulated by altering the choice incentive parameter): Because learning effects of DA in the model result from activation of D1 or D2 MSNs, the learning asymmetry persists with DA blockade due to their direct stimulation. Importantly, this is not simply a null effect: the model does predict that DA blockade reduces the absolute number of actions selected, even without changing the relative preference between the actions; both of these effects accord well with the observations of Kravitz et al. (2012; Figure 5).

Finally, Kravitz et al. (2012) reported that learning induced by D2 MSN stimulation was less robust and rapidly extinguished relative to D1 MSN stimulation. However we show here that their full pattern of results are obtained in the model without assuming any asymmetry in the robustness of learning per se. Specifically, because D2 stimulation induces avoidance, it by definition reduces the number of actions selected, and hence the number of training trials, thereby leading to weaker accumulated *N* weights and faster extinction. Indeed, D2 MSN stimulation was associated with approximately 150 actions compared with 300 actions for dMSN stimulation. Figure 6 shows that the model simulating equal effects of D1 and D2 MSN stimulation on learning reproduces both these effects of number of actions taken and, accordingly, differential apparent effects on extinction.

Tai et al. (2012) stimulated striatal MSNs expressing D1 or D2 receptors in rodents performing a reversal learning experiment using standard primary reinforcement. Optogenetic stimulation was applied at the time of choice, rather than learning, selectively enhancing activity in the direct or indirect pathway MSNs, selectively for those MSNs that correspond to only one of the available actions (i.e., action choices were left vs. right responses and stimulation was applied unilaterally). D1 versus D2 stimulation differentially impacted choices: stimulation of D1 receptors increased choice of the corresponding action, whereas stimulation of D2 receptors decreased it. Interestingly, this change in preference was not categorical (i.e., it did not just induce a motor action to left or right) but was dependent on reward history (top line of Figure 7, middle right).

To model this, we assumed the normal OpAL learning procedure as described in the methods, without optogenetic interference, but hypothesized that optogenetic stimulation affected the choice policy, selectively influencing the actor weights associated with

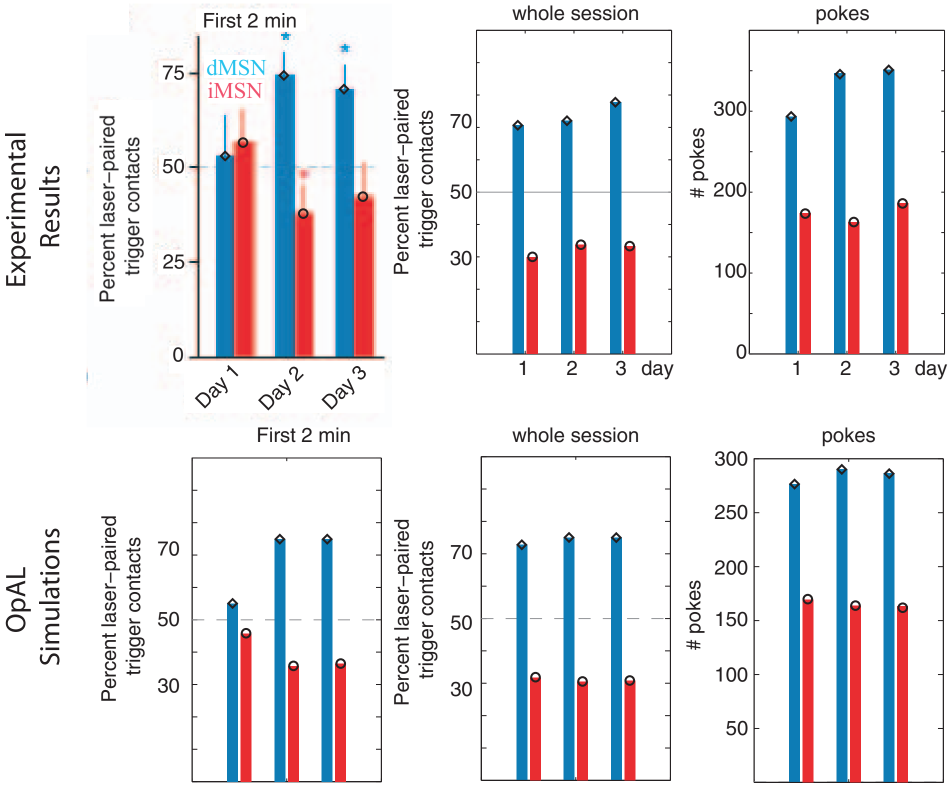

Figure 4. Optogenetic and learning. Top line: experimental results (left graph is reproduced from Kravitz, Tye, & Kreitzer, 2012; middle and right graphs are plotted from result tables in Kravitz et al.). Percent laser-paired trigger contacts refers to the proportion of trials that the animal selects the action which triggers optogenetic laser stimulation. Direct medium spiny neuron (dMSN) stimulation (bars with diamonds on top; blue bars in the online article) acted to increase likelihood of repeating the same action, whereas indirect MSN (iMSN) stimulation (bars with circles on top; red bars online) induced avoidance of the action. Bottom line: opponent actor learning (OpAL) simulations. Left: proportion of laser-paired trigger contacts in the first 2 min of the session. Middle: proportion of laser-paired trigger contacts during the whole session. Right: total number of actions (laser-paired and non-laser-paired) during the session. An asterisk represents a significant difference from chance (50), based on an alpha of .05. Top left graph adapted from "Distinct Roles for Direct and Indirect Pathway Striatal Neurons in Reinforcement," by A. V. Kravitz, L. D. Tye, and A. C. Kreitzer, 2012, *Nature Neuroscience, 15,* p. 816. Copyright 2012 by Macmillan. See the online article for the color version of this figure.

one of the actions (left or right) at the time of choice, through the increase of β_G weight in D1-MSN stimulation, and of β_N weight in D2-MSN stimulation. Parameters were symmetrical learning rates $\alpha_G = \alpha_N = \alpha_C = 0.12$, $\phi = 0.15$ and $\varepsilon = 0.5$. In absence of stimulation, softmax choice parameters were symmetrical: $\beta_G = \beta_N = 20$. Optogenetic stimulation was simulated as a 20% increase in the corresponding β weight ($\rho = -0.2$). This allows us to account for observed results (see Figure 7, bottom line): Simulations showed that stimulation produced a reward-history dependent bias toward stimulated side (D1 stimulation) or away from it (D2 stimulation).

Probabilistic Selection Task

Next we examine how model dynamics play out to explain differences in choice proportions in tasks empirically known to be sensitive to dopaminergic manipulations. Here we report simulations with a simplified and generalized version of the probabilistic selection task (Frank, Moustafa, et al., 2007; Frank et al., 2004; see Figure 3), but the same results hold with the empirical version of the task.

In this version, on each trial, the model is presented with a choice between two options. On some trials it is presented with a choice between A or B, where A is the probabilistically *most rewarding* option and B the *most punishing* option. On other trials it is presented a choice between options M_1 and M_2, which each have neutral values:

$$p(r = 1 \mid \text{choice is A}) = 1 - p(r = 0 \mid \text{choice is A}) = p > 0.5 \quad (9)$$

$$p(r = 1 \mid \text{choice is B}) = 1 - p(r = 0 \mid \text{choice is B}) = 1 - p < 0.5 \quad (10)$$

$$p(r = 1 \mid \text{choice is } M_1 \text{ or } M_2) = 1 - p(r = 0 \mid \text{choice is } M_1 \text{ or } M_2) = 0.5 \quad (11)$$

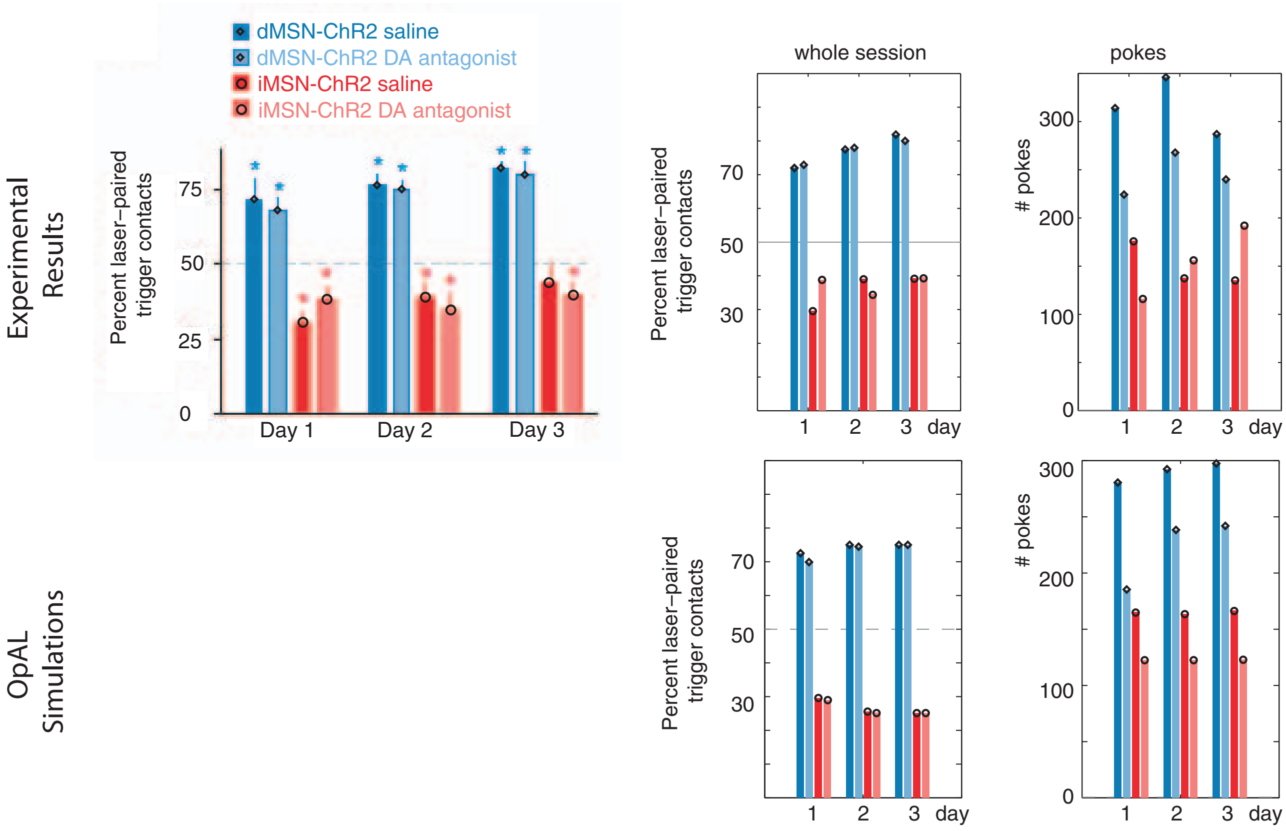

Figure 5. Optogenetic effects and dopamine (DA) blockade. Top line: experimental results (left graph adapted from Kravitz, Tye, & Kreitzer, 2012; middle and right graphs plotted from result tables in Kravitz et al.), showing similar effects on relative preference due to optogenetic stimulation even in the presence of DA antagonist. Bottom line: opponent actor learning (OpAL) simulations. Both lines, left and middle graphs: proportion of laser-paired trigger contacts during the whole session. Both lines, right graphs: DA antagonists reduced the total number of actions (laser-paired and non-laser-paired) emitted during the session, despite preserving the relative bias. An asterisk represents a significant difference from chance (50), based on an alpha of .05. Top left graph reproduced from "Distinct Roles for Direct and Indirect Pathway Striatal Neurons in Reinforcement," by A. V. Kravitz, L. D. Tye, and A. C. Kreitzer, 2012, *Nature Neuroscience, 15,* p. 817. Copyright 2012 by Macmillan. See the online article for the color version of this figure.

During the learning phase, the model reliably learns to choose A over B. In a subsequent transfer phase, the model is presented with all possible novel pairings of the four choice options (e.g., A vs. M1, A vs. M2 and B vs. M1, B vs. M2). No feedback is provided so there is no further opportunity to learn; preferences thus depend on values learned for each of the individual options during the learning phase. As in the empirical task, we define Choose-A (*ChA*) performance as the probability of picking A over M, and Avoid-B performance (*AvB*) as the probability of choosing M over B. Notably, across a range of empirical studies with this task, manipulations that increase striatal dopamine enhance *ChA* and impair *AvB*, and vice versa for manipulations that decrease dopamine; for review, see Maia and Frank (2011). Note that the difference in expected value between A and M is identical to that between M and B. Thus any performance differences between Choose-A and Avoid-B constitute a $Bias = ChA - AvB$, reflecting differential sensitivity to positive versus negative outcomes. Importantly, standard reinforcement learning models should converge to the theoretical expected value and are thus expected to produce equal *ChA* and *AvB* performance, leading to zero Bias.

To examine influences of learning and incentive without confounding effects of amount of experience (differential sampling) for different options, we first investigated a random action selection policy during the learning phase, and assess preferences between all choice options in a subsequent transfer phase. We also investigated a more standard action selection policy during learning using softmax.

We first considered how biases in sensitivity to positive versus negative outcomes (as revealed in *ChA* and *AvB* choice proportions in the test phase of this task) changed as a function of either learning phase actor learning rates α_G and α_N, or test phase choice incentive parameters β_G or β_N. Simulations (left panel in Figure 8) showed that manipulating the asymmetry either between learning rates or between test βs induced biases toward better Choose A than Avoid B for $\alpha_G >$

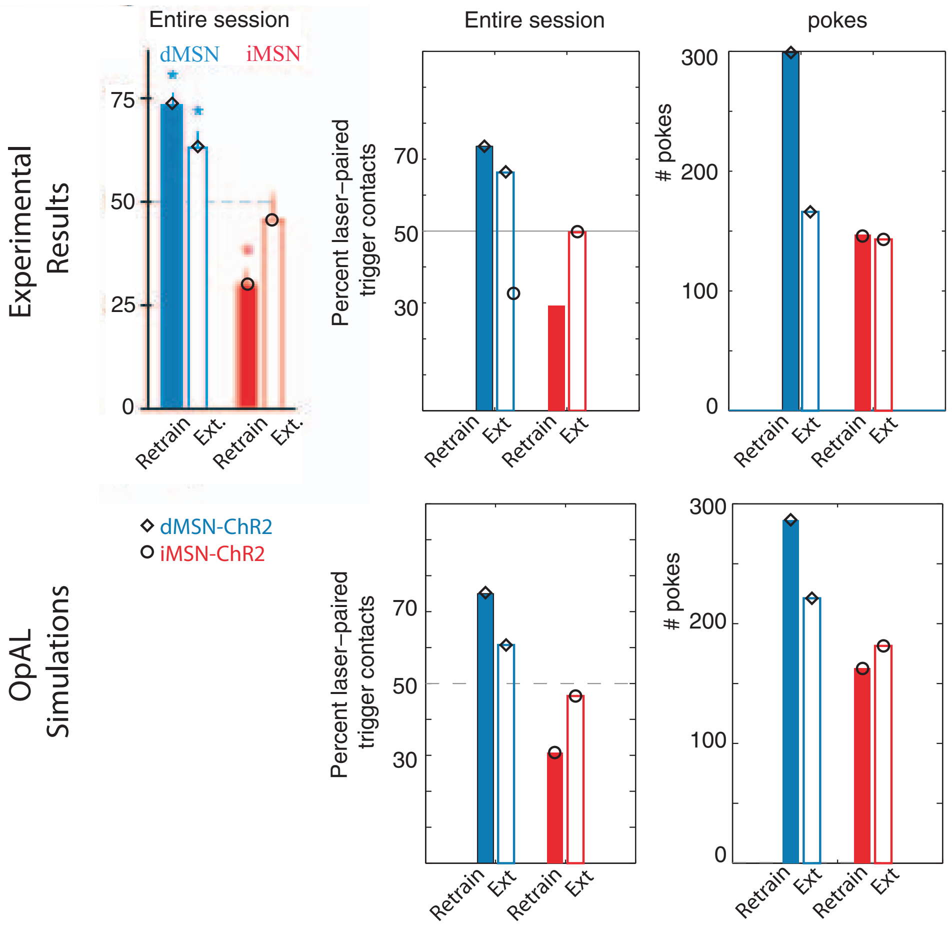

Figure 6. Optogenetic effects and extinction (Ext). Top line: experimental results (left graph reproduced from Kravitz, Tye, & Kreitzer, 2012; middle and right graphs plotted from result tables in Kravitz et al.). Findings ostensibly indicate that indirect medium spiny neuron (iMSN) stimulation is less robust and more susceptible to extinction. Bottom line: opponent actor learning (OpAL) simulations assuming equal effects of direct MSN (dMSN) and iMSN stimulation on learning reproduce the same pattern. Middle graph: proportion of laser-paired trigger contacts during the whole session. Right: total number of triggers (laser-paired and non-laser-paired) during the session. d(i)MSN-ChR2 indicates groups of mice expressing channelrhodopsin2 in d(i)MSNs. An asterisk represents a significant difference from chance (50), based on an alpha of .05. Top left graph reproduced from “Distinct Roles for Direct and Indirect Pathway Striatal Neurons in Reinforcement,” by A. V. Kravitz, L. D. Tye, and A. C. Kreitzer, 2012, *Nature Neuroscience, 15,* p. 816. Copyright 2012 by Macmillan. See the online article for the color version of this figure.

α_N or $\beta_G > \beta_N$ (and inversely for reverse asymmetries). These biases were amplified as reward probability increased/decreased for A and B, respectively.

We next fixed reward probability to that of the standard empirical task ($p(r \mid A) = 0.8$) to investigate the interaction between learning and performance parameters, by simultaneously varying α and β asymmetries. We fixed $\alpha_G + \alpha_N = 0.2$, $\beta_G + \beta_N = 2$ but parametrically altered their difference. Simulations (right panel in Figure 8) showed that bias toward ChooseA versus AvoidB increased with both $\beta_G - \beta_N$ and $\alpha_G - \alpha_N$. Interactive effects of the parameters were also visible on overall performance: when either of the set of parameters was balanced across *G* and *N*, the effect of asymmetry in the other parameter was smaller (middle column). However, when one of the parameters was strongly asymmetrical (left- and right-most columns), overall performance improved if the other parameter was asymmetrical in the same direction but dropped to chance if asymmetrical in the other direction. In other words, performance depends on motivational/dopaminergic state at the time of choice being in the same range as it was at the time of learning. Nevertheless, for more moderate learning biases, it is also possible to reverse the asymmetry in choice: when options are learned with a bias toward *N* weights, a sufficiently large asymmetry in choice incentive toward *G* weights can still result in relatively better *ChA* than *AvB* performance. These findings support the observations that motivational state at the time of choice can impact an animal’s behavior to approach an action which had been preferentially associated with negative outcomes during learning (Zhang et al., 2009).

Note that these results require the specific nonlinear update rule (including the three-factor Hebbian term) that we have introduced

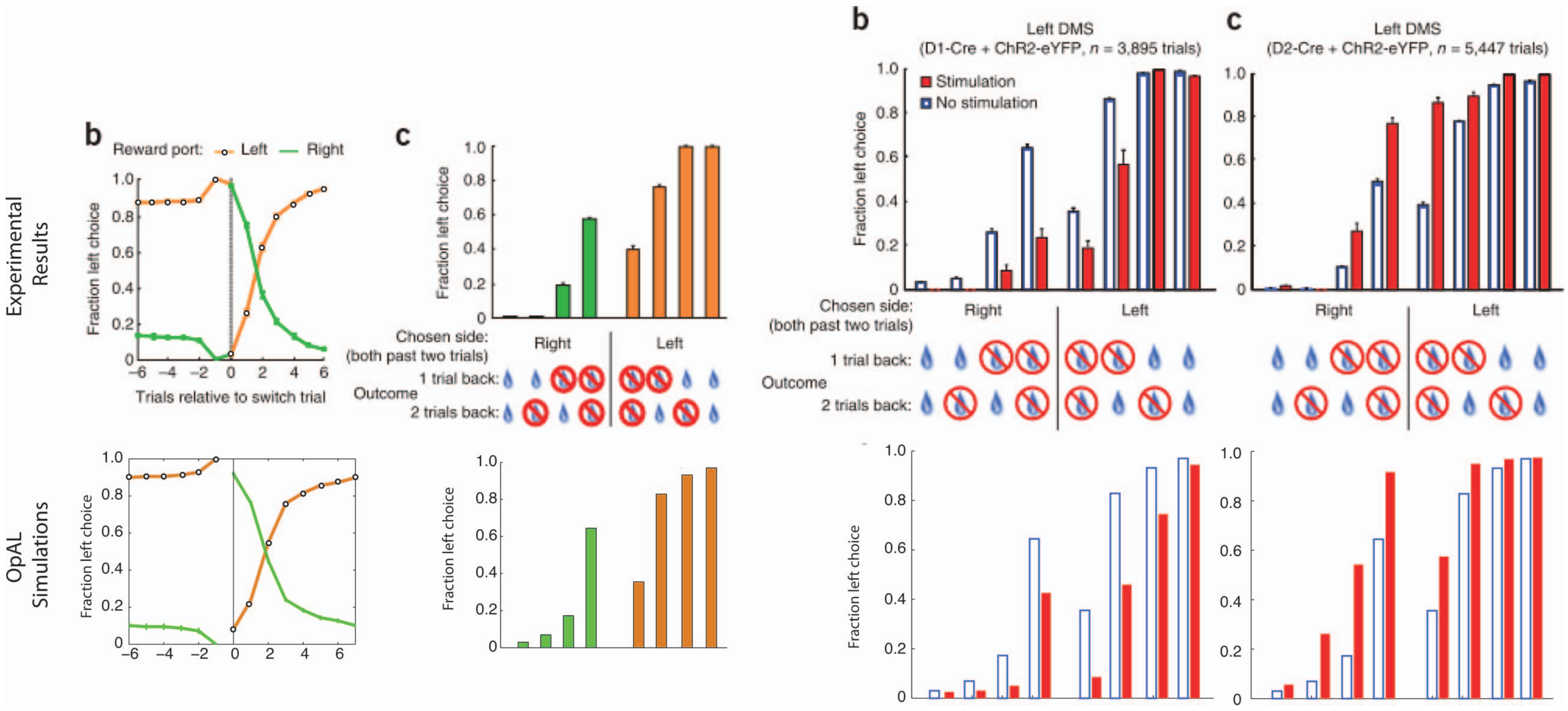

Figure 7. Optogenetic stimulation at choice time. Top line: experimental results (reproduced from Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012). Bottom line: opponent actor learning (OpAL) simulations. Left: reversal learning performance. Middle left: probability of choice for different reward history in last two trials. Middle right: proportion of left choices with and without stimulation of left D1-MSNs, for different reward history in last two trials. Right: proportion of left choices with and without stimulation of left D2-MSNs reward history in last two trials. DMS = dorsomedial striatum; MSN = medium spiny neuron. D1(2)-Cre + Chr2-eYFP indicates groups of mice manipulated for the genetic control of direct (indirect) pathway MSNs, respectively. Top row of graphs adapted from “Transient Stimulation of Distinct Subpopulations of Striatal Neurons Mimics Changes in Action Value,” by L.-H. Tai, A. M. Lee, N. Benavidez, A. Bonci, and L. Wilbrecht, 2012, *Nature Neuroscience, 15,* pp. 1282–1283. Copyright 2012 by Macmillan. See the online article for the color version of this figure.

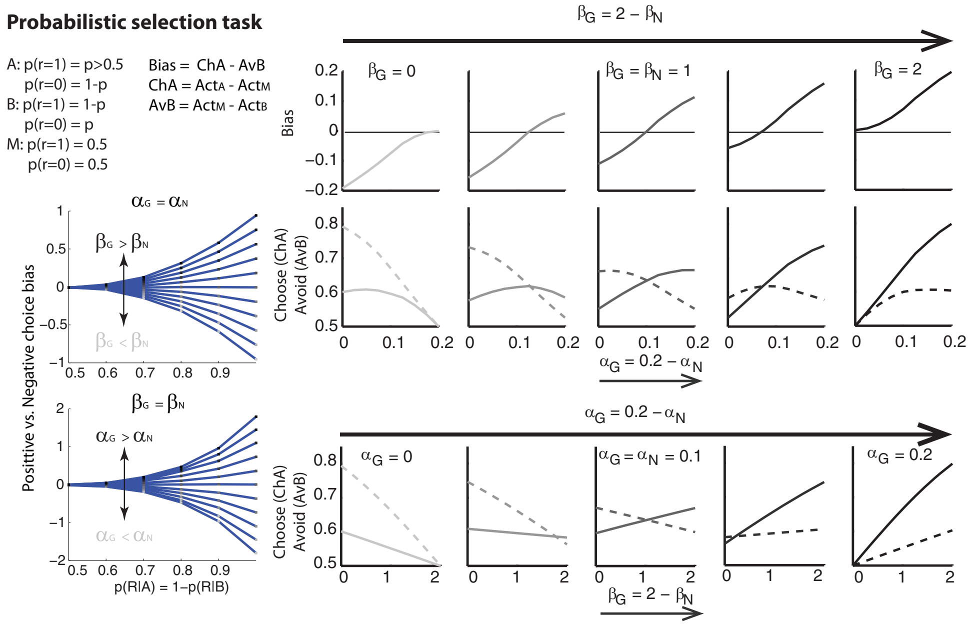

Figure 8. Simplified probabilistic selection task: relative values in opponent actor learning (OpAL). All values are final values after 100 trials, averaged over 1,000 simulations. Left: choice bias for different probabilities of reward $p(r)$. Top left: fixed $\alpha_G = \alpha_N$, varying β_G versus β_N asymmetry. Bottom left: fixed $\beta_G = \beta_N$, varying α_G versus α_N asymmetry. Both learning or incentive (performance) effects can produce a choice bias emphasizing positive or negative values. This bias increases as the most rewarding/punishing outcomes are increasingly deterministic. Right: Choose-A (full line) versus Avoid-B (dotted line), and Bias (relative difference) as a function of asymmetries in α and β parameters. Bias increases monotonically with asymmetries in either parameter type, but effects of both parameters interact: Given an asymmetry in learning (α), performance is best when the asymmetry in incentive (β) is in the same direction, i.e., when dopaminergic motivational state at the time of choice is similar to that at the time of learning. Even so, for intermediate levels of asymmetry, it is possible to exhibit greater learning in one system but to express greater influence of the other during choice, as in some experiments (Zhang, Berridge, Tindell, Smith, & Aldridge, 2009). Horizontal black lines in the top line of graphs show simulations for a version of the model without the Hebbian nonlinear term in the actor weight updates: This model cannot account for differential sensitivity of Choose A or avoid B, as can be seen by null bias across all parameters, due to symmetrical representation of positive and negative values in both *G* and *N* weights. See the online article for the color version of this figure.

for OpAL. Indeed, the gray curves in Figure 8 show simulations of a simplified model stripped of the multiplicative modulation by *G* or *N* values in the update equations. These simulations show that no bias is observed with any combination of asymmetries of learning rates or β parameters. Without the multiplicative update, *G* and *N* weights evolve symmetrically, with no preferential differentiation among positive or negative values, and as such are linear combinations of true expected value, leading to equal Choose-A and Avoid-B performance (see Appendix, supplemental simulations in Figure A2).

Dopaminergic manipulations in the probabilistic selection task have been shown repeatedly (Frank, Moustafa, et al., 2007; Frank & O'Reilly, 2006; Frank et al., 2004; Jocham et al., 2011; Shiner et al., 2012; Smittenaar et al., 2012) to induce changes in Choose-A versus Avoid-B bias, although it has not been clearly disentangled whether this was due to learning effects or performance effects. As noted earlier, some studies show effects of dopamine medication in which Choose-A performance is improved even when the design was such that medications could have only affected test performance rather than learning (Shiner et al., 2012; Smittenaar et al., 2012). However, other experiments in which dopamine was modulated during both learning and test showed greater effect on choice asymmetry than these showing effects at test alone. Further, imaging studies showed that the effect of dopaminergic manipulations on Choose-A performance is correlated with the extent to which it boosts reward prediction error signaling during learning (Jocham et al., 2011; Ott, Ullsperger, Jocham, Neumann, & Klein, 2011). The above simulations suggest that both learning and incentive motivational effects could account for the results either separately or jointly. They are also consistent with the fact that dopamine modulation effects should be parametric, as observed in genetic studies (Frank, Moustafa, et al., 2007).

Choice bias in the probabilistic selection task have usually been modeled with classic reinforcement learning models that include asymmetric learning rates for positive versus negative errors (Doll et al., 2011; Frank, Moustafa, et al., 2007). We show in further simulations in the Appendix (Figure A1) and in the discussion that while such models can indeed account for some bias effects, they cannot account for the variety of data that OpAL can.

Our model can thus account for previously observed effects of dopamine in the probabilistic selection task, both in the learning stage and in the subsequent performance stage. However, it also makes additional predictions, in an experimental design where dopamine would be manipulated separately during the learning phase and during the test phase. As seen in Figure 8, middle, OpAL predicts that the bias is exaggerated if both learning rates and β parameters are adjusted. Moreover, it predicts that overall performance declines precipitously if the learning and testing phase are performed in different dopaminergic states. For example, near chance performance is predicted given high dopamine during learning (strong $\alpha_G > \alpha_N$ asymmetry) but low dopamine during test (strong $\beta_N > \beta_G$ asymmetry; see Figure 8, right group, middle left plot).

Motivation and Incentive Effects on Effort-Based Decision Making

Perhaps the most clear example of dopamine on motivational incentive comes from tasks that manipulate the amount of effort an animal (or human) has to exert to attain a reward (Cousins & Salamone, 1994; Floresco, Tse, & Ghods-Sharifi, 2008; Salamone et al., 2005; Treadway et al., 2012). Here we consider the paradigm in which humans or animals need to press a single lever a number of times T to obtain a reward. In such studies, dopamine modulations strongly influence the degree of effort exerted, such that the tendency for animals to work harder for higher potential rewards is proportional to striatal DA: It is enhanced with DA elevations and suppressed with DA depletions. These findings are not easily accounted for by learning theories because manipulations are conducted after the animal or human has learned the effort cost and reward benefit contingencies. Nevertheless, this procedure mainly reflects the tradition from which effort-based decision making has been studied, and there is no principled reason why dopamine manipulations could not be conducted during learning itself. We thus consider potential roles both of learning and of motivational incentive and their interaction.

In a first simulation, we parametrically varied the number of actions required to attain a reward, indexed by the reward probability associated with a single choice. We used a fixed learning period (100 lever presses) to acquire the contingencies, and symmetrical learning rates but varied the balance ρ between β_G and β_N during choice (see supplemental methods in Appendix). Figure 9 shows that increasing ρ, thus $\beta_G > \beta_N$, leads to a greater probability of selecting the option. This effect interacts with the amount of effort required to obtain a reward: for high effort (e.g., $p(r) = 0.1$), the effect of changing ρ further becomes very apparent once there is any bias for $\beta_G > \beta_N$ (green curves), whereas the same change in ρ for the opposite asymmetry has little effect. Conversely, for low effort (e.g., $p(r) = 0.9$), the effect of changing ρ by the same amount has a far greater effect given the opposite asymmetry $\beta_N > \beta_G$ (red curves).[4]

In a second set of simulations, we directly modeled dopamine's influence on the D2 pathway, in accordance with evidence that manipulation of D2 receptors and of adenosine A2A receptors (which are colocalized on the same NoGo neurons) predictably modulates the cost of effort. Specifically, D2 blockade effectively enhances the cost of effort, whereas A2A blockade, by having opposing effects on neuronal excitability, counteracts this effect (Farrar et al., 2010, 2008; Mingote et al., 2008; Nunes et al., 2010). We explored potential effects on learning the effort cost (varying $\alpha_N = 0.1$ or 0.125) and on the expression of this cost ($\beta_N = 1$ or 1.5), while keeping G parameters fixed ($\alpha_G = 0.1$, $\beta_G = 1$). Simulations revealed that D2 blockade either during learning (gray vs. black lines), or during performance (green circles vs. red squares) reduce the effective actor weight *Act* of the option, and hence the probability of engaging in the effort necessary to select the option. Notably, these D2 effects increase with the cost of the option (number of presses required).

These simulations also provide a novel, testable prediction. In addition to the main effects of effort cost on action engagement, and the effects of both learning and incentive parameters, we also observe interactions between each pair of factors, as well as a three-way interaction (all $ps < .01$). Learning and performance effects of D2 manipulation are stronger the more effort is required (as evidenced by the increasing distance between the curves), and the combination of both manipulations during learning and choice amplifies the effect. Moreover, contrasting the two effects individually, the effects of choice incentive are stronger than learning effects when effort cost is relatively low, but for high costs, this can be reversed (compare −/+ and +/− conditions for 2–4 vs. 16–20 lever presses). This is because with high effort, the increased frequency of negative reward prediction errors (lack of reward for most presses) accumulates over time, due to the multiplicative influence of N weights on updates, resulting in stronger learning effects. Thus, the triple interaction is a specific prediction of the OpAL model, not predicted without the Hebbian term (see Appendix, Figure A2).

In addition to providing novel, testable predictions, we show more directly that the OpAL framework can account for existing data by simulating it on common effort protocols, including those that manipulate the number of lever presses required to obtain a reward with and without dopamine blockade (Aberman & Salamone, 1999; Niv et al., 2007), and those that add a barrier that a rat has to climb over to obtain a larger reward in a T-maze. In both cases, the model learns choice contingencies in the intact state, modeled with $\beta_G = \beta_N$ and $\alpha_G = \alpha_N$, then is tested in extinction (no learning), either in the intact state or with dopamine blockade (modeled as an asymmetry in parameters $\beta_G < \beta_N$).

Specifically, to model the environment of the lever-pressing effort task, we assume a single state and choice, that is selected or not according to Equation 6. When selection occurs, feedback is modeled with a Bernoulli probability $p(r) = 1/T$. We allow the model to learn these contingencies and manipulate dopamine effects during this learning through actor learning rates α_G, α_N. We then manipulate motivational incentive following learning in extinction, by modulating the $\beta_{G,N}$ parameters: we varied the balance between β_G and β_N during choice, by setting $\beta_G = \beta * (1 + \rho)$, $\beta_N = \beta * (1 - \rho)$, with $\beta = 1$. Parameter $-1 < \rho < 1$ represents the normalized difference between β_G and β_N.

To apply this specifically to data from Aberman and Salamone (1999), we used the following parameters: $\beta = 10$, $\alpha = .02$, $\rho = 0$, $\rho_{DAblock} = -0.9$. The model learned to choose between two

[4] Note that this cannot be reduced solely to a shift in a single softmax curve.

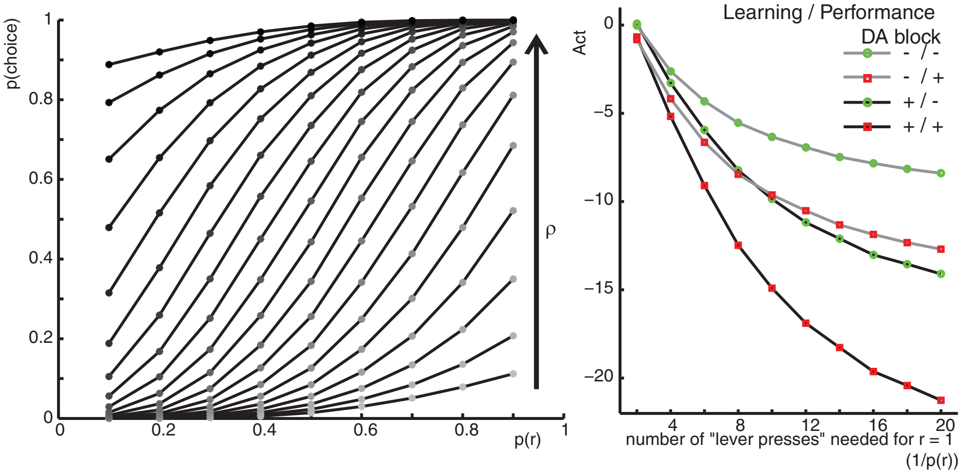

Figure 9. Effort tasks: manipulation of β asymmetry in single forced choice lever pressing case. All values are final values after 100 lever presses, averaged over 1,000 simulations. ρ is normalized difference $\frac{\beta_G-\beta_N}{\beta_G+\beta_N}$. Left: Increasing asymmetry toward *G* system enhances willingness to engage in effort. Darker colors indicate asymmetry toward *G* vs. *N* ($\rho < 0$). Right: With α_G and β_G fixed, simulation of D2 receptor blockade during either learning (higher α_N) or subsequent choice (higher β_N), or both. Gray versus black lines indicate control versus drug during learning, green circles versus red squares indicate control versus drug during performance. Both manipulations produce decreased effort, especially when combined (+/+), and these effects are magnified with increased effort cost. Learning effects alone (+/−) are greater than performance effects alone (−/+) for high effort cost; this pattern reverses for decreased effort cost. DA = dopamine. See the online article for the color version of this figure.

options: press the lever or do nothing. If press was chosen, the probability of reward $r = 1$ was $1/FR$, with fixed ratio $FR = \{1, 4, 16, 64\}$. The reaction time was modeled as $RT = 0.5 + 1.5/(1 + exp(Act))$. When reward was obtained, we assumed a fixed eating time ($\tau_I = 6$). Simulations were ran for two 30-min equivalent sessions, then choice results plotted from the third posttraining session (following the fact that animals are extensively trained prior to performance in these paradigms).

Thus, in the lever press task, OpAL simulations reproduce the basic pattern commonly observed for fixed ratio schedules: an increase in lever presses as the schedule demands increase but where dopamine blockade preferentially decreases lever pressing as effort increases (Figure 10A). The model suggests that this pattern results from the fact that the high effort condition is associated with greater cost that is traded off against the reward benefit, where the cost is exaggerated with dopamine blockade.

In the T-maze task, animals learn to choose the arm providing most food pellets (4 vs. 0, or 4 vs. 2). Intact animals continue to choose the arm with four pellets even when a barrier is added such that they have to climb over it. However, dopamine blockade induces them to stop choosing the most rewarding arm in the 4 versus 2 case, but not in the 4 versus 0 (Cousins, Atherton, Turner, & Salamone, 1996; Figure 10B). The observation that dopamine blockade preserves choice of the high effort option in the 4 versus 0 case suggests that effort is not coded distinctly from reward value (i.e., the animals are able to climb the barrier if they want to) but that they are rather performing a cost-benefit analysis, i.e., that the cost of climbing the barrier is compared in some currency to the benefit of the reward.

We modeled the T-maze with barrier task in Cousins et al. (1996) by assuming a single state, and two possible actions (left arm, right arm). The left arm choice deterministically provided $r = 4$ pellets, while the right arm choice deterministically provided either $r = 0$ or 2. The model was trained for 100 trials and learned to robustly choose the left arm. To model the barrier effort, we separately trained the model in an environment where it had to pick the action [climb the barrier] and assumed this action led to a cost of $c = -1$ for 100 trials. Thus the model developed *G* and *N* weights for that action. The model was then tested on the combination of the T-maze with barrier: animals had to climb a barrier in the left arm in order to access pellets. We assumed that the choice was made between {left arm and barrier} and {right arm} so that the actor weights considered for the left choice were the sum of the learned left arm and barrier weights. OpAL parameters were $\alpha s = 0.1$, $\beta = 3.5$, $\rho = 0$. Dopamine blockade was modeled with $\rho = -0.55$, leading to $\beta_G < \beta_N$. Results are averaged over 1,000 simulations. Figure 10B (right) shows that OpAL simulations reproduce the behavioral pattern quantitatively.

We have now shown that this model accounts for the two main classes of results of dopamine effects on learning and motivation. Differential sensitivities to gains versus losses in reinforcement learning experiments have heretofore been attributed primarily to learning effects, which the model can capture. But the model also shows that choice incentive/performance effects can also contribute to those findings. Symmetrically, dopamine effects on effort-based decision making have largely been studied in the context of motivational incentive models (performance effects), which our model can capture but shows that learning effects could also contribute if dopamine is manipulated during that period.

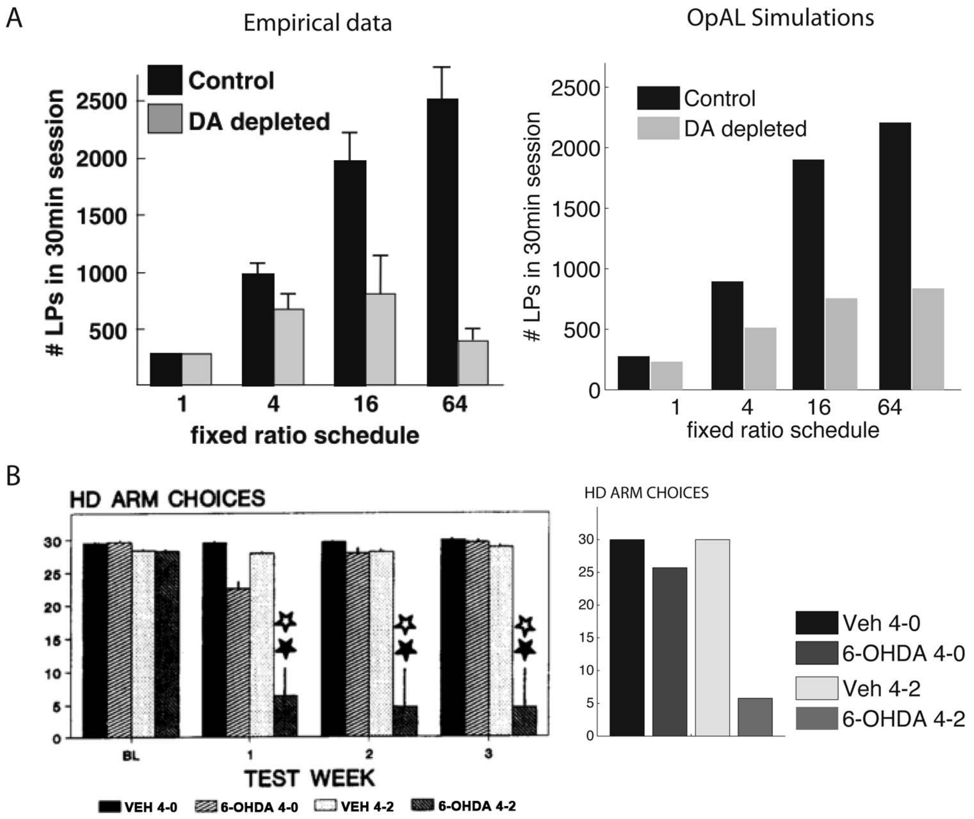

Figure 10. Effort tasks. A. Fixed ratio lever press task. Left: data from lever press task (from Aberman & Salamone, 1999). Right: opponent actor learning (OpAL) simulations of the task. As in the data, the number of lever presses increases with fixed ratio schedule, but dopamine depletion decreases the number of lever presses as effort demands increase. B. T-maze with barrier task: Experimental data from Cousins, Atherton, Turner, and Salamone (1996) on the left, model simulations on the right. Healthy rodents and intact models prefer the arm with the largest reward despite the increased effort. Dopamine depletion reverses this preference, but only for the 4 versus 2 pellet case, without impacting choice in the 4 versus 0 pellet case. LP = lever press; DA = dopamine; HD = high density food; BL = baseline; Veh = group injected with vehicle; 6-OHDA = group injected with 6-hydroxydopamine. Full star indicates significant difference from Veh condition, open star indicates significant difference between 4-2 and 4-0 condition. Error bars are standard error of the mean. Left graph in Panel A reproduced from "Tonic Dopamine: Opportunity Costs and the Control of Response Vigor," by Y. Niv, N. D. Daw, D. Joel, and P. Dayan, 2007, *Psychopharmacology, 191,* p. 512. Copyright 2007 by Springer. Left graph in Panel B reproduced from "Nucleus Accumbens Dopamine Depletions Alter Relative Response Allocation in a T-Maze Cost/Benefit Task," by M. Cousins, A. Atherton, L. Turner, and J. Salamone, 1996, *Behavioural Brain Research, 74,* p. 192. Copyright 1996 by Elsevier.

Both of these sets of simulations further suggested potential interactive effects of learning and performance, as novel predictions. To more concretely link these predictions to empirical data, we next turn to motor skill learning experiments that have convincingly shown these interactions.

Learning and Performance Interactive Effects on Motor Skills

Striatal dopamine has long been implicated in motor performance, but its role in learning, and particularly the impact of DA depletion on aberrant learning, has only recently been appreciated. In Beeler et al. (2012), the authors administered dopamine blockade to rodents performing the accelerating rotarod task, a motor skill learning task where rodents are put on a rod that turns with accelerating speed: The animal integrates visual and proprioceptive feedback to walk forward at the correct rate to avoid falling. They showed interactive effects on learning and performance, which were captured by a neural network model of basal ganglia. In particular, dopamine blockade either during first exposure to the rotarod or after having learned the task led to very poor performance. While superficially these findings could be attributed to performance deficits, further results showed that it also induced aberrant learning. Indeed, after drug washout, learning was significantly slower than for naive animals. Similarly, after having learned the task in an intact state, D2 blockade in particular resulted in progressive decline in skill performance. Synaptic plasticity studies showed that D2 blockade induced potentiation of corticostriatal synapses onto D2 MSNs. These same mechanisms have been proposed to account for the progressive development of Parkinsonian symptoms given repeated administration of low dose D2 blockers in catalepsy experiments (Wiecki et al., 2009). We tested here if our largely simplified version of the neural network reinforcement learning mechanism could account for the effects on motor skill learning and performance.

Our OpAL simulations made similar assumptions as previous neural network simulations modeling this task: We assumed four states and four motor actions (simplistically, corresponding to which paw to move). Moreover, the correct action needed to be taken rapidly enough, otherwise the animal would fall off the rod. We thus assumed that correct action choice leads to reward ($r = 1$) with a probability dependent on reaction time:

$$p(r \mid correct) = 0.1 + .8/(1 + \exp(\beta_G * G(s, a) - \beta_N * N(s, a))),$$

but a punishment ($r = 0$) otherwise; whereas incorrect actions always lead to a punishment ($r = 0$). Parameters used are $\beta = 3$, $\alpha_{G/N} = 0.1$, $\alpha_C = 0.05$, $\rho_{drug} = -0.75$, $\rho_{control} = 0.5$, with $\beta_G = \beta * (1 + \rho)$, $\beta_N = \beta * (1 - \rho)$.

Figure 11 shows the main simulation results of the rotarod task. Performance effects of DA blockade are modeled here by setting $\beta_G < \beta_N$ (while the nontreated condition has $\beta_G > \beta_N$), with all other parameters kept fixed (in particular, $\alpha_G = \alpha_N$). In a first simulation (Figure 11A), DA blockade at first encounter with the task (colored part) accentuates NoGo activity, leading to weak actor weights *Act* (bottom right), thus slowed action selection, even for actions that would have otherwise been correct. Thus, this performance effect means that even correct actions are rarely rewarded, leading not only to a lack of learning (no increase in performance, top graph) but also to *aberrant* learning: a decrease in *G* weights (bottom left) and increase in *N* weights (bottom middle), even for correct actions. This is revealed by subsequent exposure to the task in the intact dopaminergic state (white region). Learning then proceeds, with the model correctly learning *G* and *N* weights for correct and incorrect actions (full black and gray lines), but slower than that in the naive case (dotted lines).

In a second set of simulations (Figure 11B), the model first learns the task normally. Exposure to DA blockade after establishment of the skill leads to a rapid drop in performance but also to aberrant learning: *G* weights decrease and *N* weights increase for both correct and incorrect actions, leading to progressive decline in performance. This again makes subsequent relearning of the task in the absence of blockade slower, for the same reasons as the previous experiment.

We thus showed that the OpAL model can account for interactive effects of learning and performance in the rotarod task, including aberrant learning due to performance effects of dopamine blockade. This pattern of results is again dependent on having the multiplicative update rule in OpAL.

Instruction Bias

Finally, we considered a higher level cognitive interaction between performance and learning: how top-down, rule-guided instruction in humans can affect performance and bias learning. We simulated the instructed probabilistic selection task, a variant of the probabilistic selection task in which one of the six stimuli is (rightly or wrongly) shown to the subject prior to learning of task contingencies, framed as a hint that this option is likely to be a good choice.

Specifically, this task builds on the basic probabilistic selection task (Frank, Moustafa, et al., 2007): During an initial learning phase, subjects learned to pick the most rewarding option for three pairs of stimuli (pair AB with $p(r \mid A) = 0.8 = 1 - p(r \mid B)$, pair CD with $p(R \mid C) = 0.7 = 1 - p(R \mid D)$, and pair EF with $p(R \mid E) = 0.6 = 1 - p(R \mid F)$); then, during the transfer phase, all possible pairs of stimuli are presented, but there was no feedback following choice. Bias was measured again during the test phase as $Bias = ChA - AvB$, with Choose-A (*ChA*) defined as performance on choosing A over lower valued stimuli C, D, E, or F (which have on average value of 0.5 and thus correspond to the simplified version of comparing A to M); and Avoid-B (*AvB*) as performance on avoiding Stimulus B in favor of these same more neutral choices. In the instruction bias version of this experiment (Doll et al., 2011; Doll, Jacobs, Sanfey, & Frank, 2009), prior to the learning phase one of the six stimuli was shown randomly and instructed the subjects that it was likely to be *good*, truthfully or not.

Experimental results (Doll et al., 2009) showed that subjects initially select this instructed option, but when instructions were misleading they eventually learned to avoid it during the training phase. Nevertheless, transfer phase choices indicated that the learned value of this instructed stimulus was inflated relative to uninstructed options of the same objective value. Indeed, model fits suggested that a confirmation bias could account for the findings, where during the training phase outcomes that were consistent with the instructions were amplified, and inconsistent ones discounted, leading to an inflation of objective value.

Genetic results (Doll et al., 2011) showed that a DARPP32 polymorphism (a dopaminergic genetic variant influencing plasticity in opposite directions in D1 and D2 pathways) was linked both to asymmetries in Choose-A relative to Avoid-B performance, whereas DRD2 polymorphism related to D2 receptor function was related to Avoid-B performance, as had been previously reported. Moreover, in the instructed version, these genetic variants were, respectively, predictive of the tendency to amplify the values of instruction-consistent outcomes and to discount the negative outcomes. Behaviorally, this amounted to better ability to choose the instructed stimulus when it was appropriate to do so (Choose-I) and worse ability to avoid it (Avoid-I) when paired with more valued options. Thus, genes associated with basic uninstructed reinforcement learning are also predictive of the extent to which learning is subject to confirmation bias, suggesting that this bias in the instructed experiment arises not from a separate mechanism (e.g., higher level strategy) but from a modulation of those same basic RL mechanisms. We thus test our OpAL model to see whether its pure reinforcement learning mechanisms can account for this set of data.

Here we attempted to account for these findings in OpAL by modeling initial instruction about a given stimulus being *good*, by simply boosting its initial *G* weight and depressing its initial *N* weight by a fixed value (0.3). Simulations showed that OpAL model could account for this array of results, suggesting that DARPP-32 modulates learning asymmetry but that DRD2 modulates choice incentive. First, Figure 12 shows that the model learns away from a wrongly instructed stimulus, without completely overcoming the initial bias (compare cyan instructed to gray un-

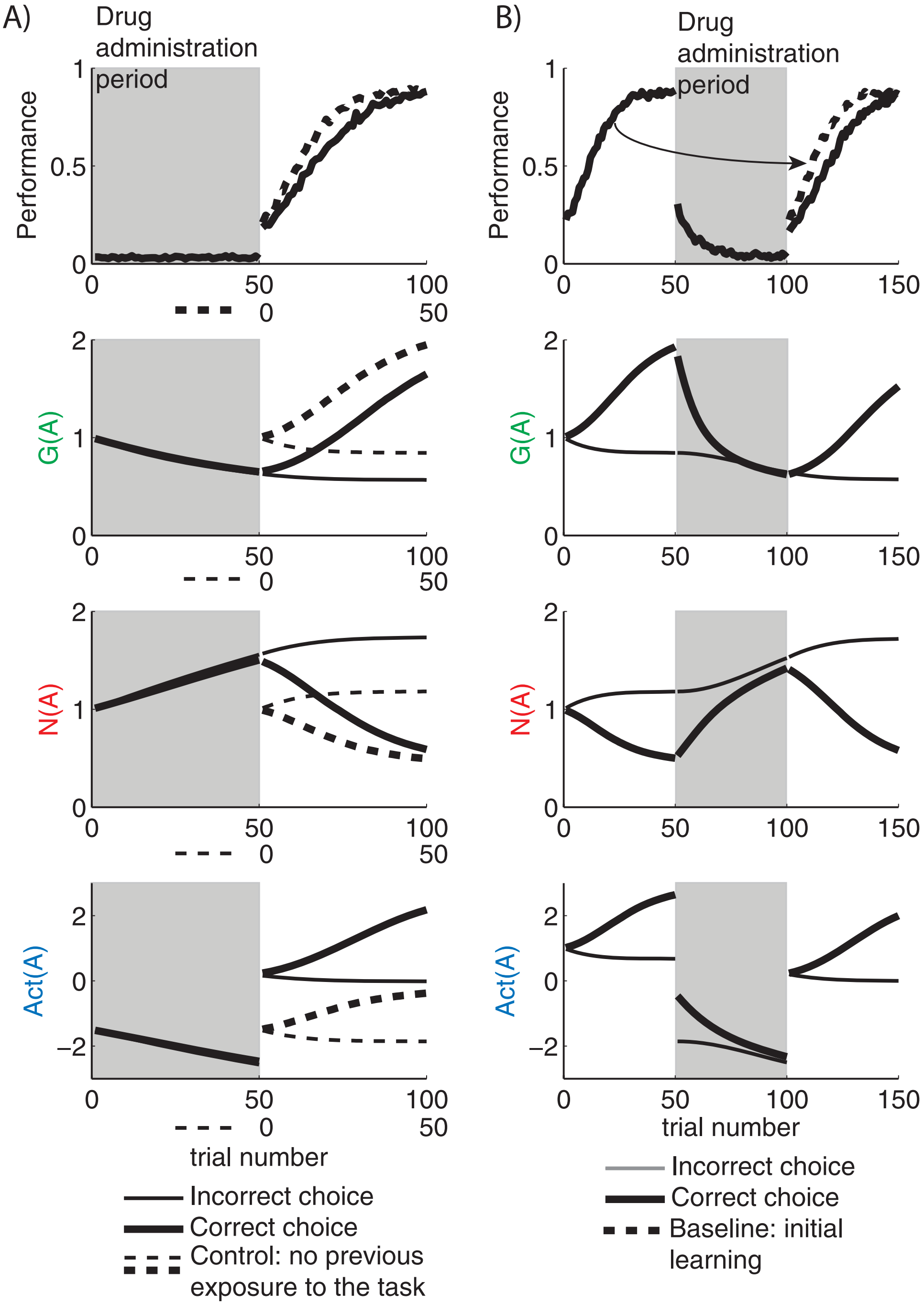

Figure 11. Rotarod task simulations. Dopamine (DA) blockade effects on performance are modeled by setting $\rho = -0.75$ instead of $+0.5$, with all other parameters fixed. Periods of drug administration are indicated by gray background. A. Top: performance (proportion of rewarded trials) over time for drug followed by intact (full line) or for intact control without previous exposure to the task (dotted line). Bottom: G and N weights and actor values *Act* for model parameters. Performing is impeded during first presentation of the task with drug and leads to avoidance learning of both correct and incorrect actions. This provokes slower learning after the drug is removed, compared to controls, as seen in empirical studies. B. In second set of simulations, drug is administered after the task is learned, leading to a rapid drop in performance that further degrades with time. In the third phase, without drug, relearning is again significantly slowed compared to initial learning (dotted line). See the online article for the color version of this figure.

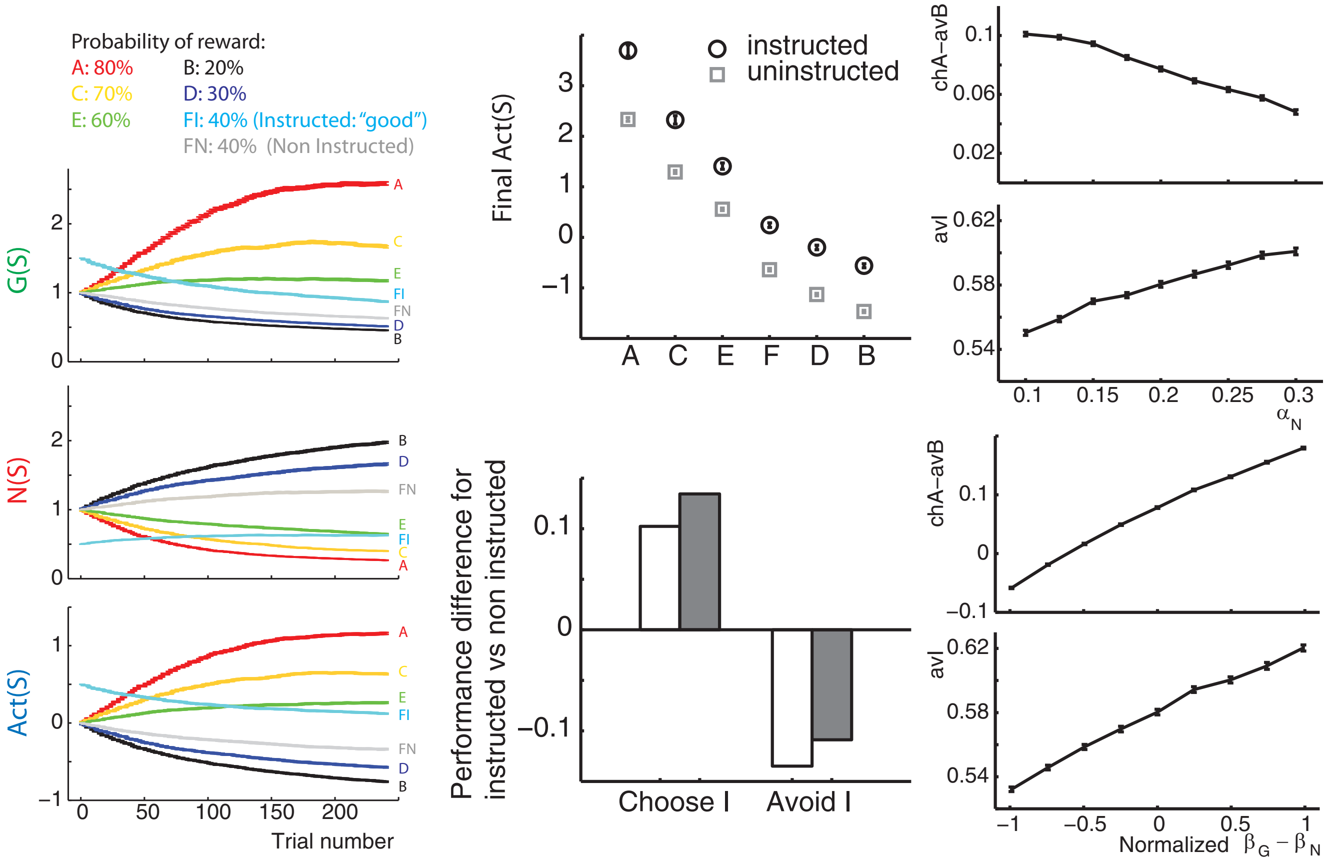

Figure 12. Instructed probabilistic selection task. Left column of graphs: Model values for different stimuli over time: *G* and *N* weights, Actor values *Act*. The line labeled FI (light blue in the online version of the figure) gives an example of misleadingly instructed *F* stimulus (objective reward probability is 40%) as *good*, compared to the line labeled FN (gray), for uninstructed *F*. Note that the asymmetry in actor weights is stronger in *G* than *N*, compared to Figure 2. This is because sampling is unequal: Over time, the model learns to choose A, C, and E more often and thus learns more about them. Since these are positively valenced options, the effects are more visible in *G* weights. Note that *Act*(*F*) ends up below *Act*(*E*), showing that the instruction bias becomes unlearned during training but that the prior initialization persists, as it nevertheless does not catch up with its uninstructed version value. Middle top: Final *Act* values for instructed (black circles) and uninstructed (gray squares) stimuli. Error bars indicate standard error of the mean. Middle bottom: Effect of instruction on test phase, showing relatively increased choice of instructed versus noninstructed stimuli (higher Choose-I and lower Avoid-I) for both accurate (white) and inaccurate (gray) instructions. Right column of graphs: Parametric gene effects: Increasing α_N leads to relatively worse Choose-A (ChA), compared to Avoid-B (avB), but simultaneously improves Avoid-I (avI) performance, as observed for the DARPP-32 polymorphism. Increasing β asymmetry ($\beta_N < \beta_G$) decreases Avoid-B performance but increases Avoid-I, as observed for the DRD2 polymorphism. See the online article for the color version of this figure.

instructed, for OpAL *Act* values). More generally, the top middle graph shows an overall boosting of asymptotic *Act*(*s*) when this stimulus has been instructed as good, as was observed experimentally. Simulations also reproduce test performance Choose-I versus Avoid-I, showing better performance at choosing a stimulus over a statistically worse stimulus when it has been instructed as *good* and increasingly so when this instruction was misleading (gray bar). Conversely, subjects have a harder time avoiding to choose a stimulus over a statistically better stimulus when it has been instructed as *good*, but less so when the instruction was misleading.

Finally, simulations can reproduce genetic effects (see Figure 12 right panel). First, increase in learning parameter α_N compared to α_G leads to a decrease in the Choose-A versus Avoid-B bias by increasing No-Go learning efficiency. This also leads to increased avoid I performance by providing better unlearning of wrongly instructed stimuli. Thus, DARPP-32 effects can be accounted for by changes in asymmetry in learning parameters, as originally interpreted (Doll et al., 2011)). DRD2 effects, by contrast, are not accounted for by learning asymmetries but instead are accounted for by choice incentive effects (β asymmetry). As described above, with $\beta_G > \beta_N$, Avoid-B performance is reduced, but counterintuitively, this also increases Avoid-I performance. The reason for this pattern is that due to the prior instruction effects, the instructed stimulus I develops stronger *G* rather than *N* weights, and hence differentiating between it and higher valued *G* weights depends on relatively stronger influence of β_G. This provides an explanation for how the DRD2 gene can modulate the sensitivity to uninstructed negative outcomes in one direction, but those to instructed stimuli in the opposite direction. Note that this account diverges from the original interpretation of the role of DRD2 in terms of *learning* effects. Indeed, we suggest here that it may be better explained by *choice incentive* effects, which accounts for both standard and instruction effects we observed. It remains to be seen whether previously documented effects of DRD2 on uninstructed avoidance are due to incentive rather than learning effects.

Why Have Two Systems: Normative Analysis

We have shown that OpAL can account for a wide pattern of experimental data relating to reinforcement learning and decision making, as well as dopaminergic influence thereof. These simulations rely crucially on OpAL's structure as a dual representation mechanism, with G and N coding for strongly anticorrelated values. This relative redundancy between information in the direct and indirect pathways generates the question of why two systems, coding for negatively correlated value estimates, are necessary or beneficial. Intuitively, this system provides for added flexibility, such that whether one emphasizes distinctions between learned prospective rewards or costs can be subject to their current motivational state, i.e., the level of dopamine at the time of choice. Future work will investigate how this state can itself be optimized as a function of other variables.

Here, we focus on the ability of OpAL, even without any asymmetry in learning or choice parameters, to learn probabilistic contingencies and compare its performance to standard RL algorithms (results presented below hold for both standard actor-critic and Q-learning). Models were presented with two pairs of options to choose from, where overall reinforcement schedules were either *rich* (probabilities of reward $r = 1$ vs. 0 of 0.8 and 0.7) or *lean* (0.3 and 0.2). Over time, models should learn to pick the objectively optimal option for each pair (0.8 and 0.3).

We optimized model parameters to obtain best mean performance in picking the optimal option over 50 learning trials across 10,000 simulations (see supplemental methods in Appendix). As noted above, we constrained symmetry, with $\alpha_G = \alpha_N$ and $\beta_G = \beta_N$, to investigate the specific role of two systems even without imbalance between them. Simulations with optimized parameters showed that OpAL performed better on average than an RL model (for which parameters were also optimized) and that this was particularly true for learning about the *lean* options (Figure 13). Follow up analysis indicated that in RL, learning for the lean option was slowed compared to the rich one, because of an exploitation/exploration conflict. A mathematical derivation of this problem for RL is presented in the Appendix. Intuitively, in the 20/30 discrimination, as soon as the model begins exploiting the 30 option, it fails to learn the true value of the worse one (20), so that its estimated value remains closer to initialization (0.5), and hence closer to the 30 option. Similarly, for the 70/80 option, once the model exploits the 80 choice, its estimated value for the 70 option remains closer to 0.5, but in this case it is helpful because the effective difference between the exploited and nonexploited values is larger. Thus for the same softmax β parameter, the RL model does not discriminate between 30 and 20 as well as it does between 70 and 80. Moreover, simply increasing the β parameter does not help (indeed it was optimized), because overly high values prevent exploration to acquire true contingencies of alternative options.

How does OpAL avoid this problem to perform equivalently well for rich and lean choice discriminations? Of critical essence, OpAL includes both G and N weights into its choice function. Thus, initially, before either set of weights accumulates sufficient values to dominate the other, they contribute relatively equally. In the 20/30 case, the differences among learned G values is de-emphasized (e.g., Figure 2), and hence this half of the choice function effectively increases exploration. However, once the N weights accumulate sufficiently, they dominate, and hence the contribution of G weights is negligible, and the model can effectively exploit the 30 option. This functionality implies that exploration is dynamically regulated as the system learns to favor either G or N weights appropriately. This functionality crucially relies on the nonlinear rep-

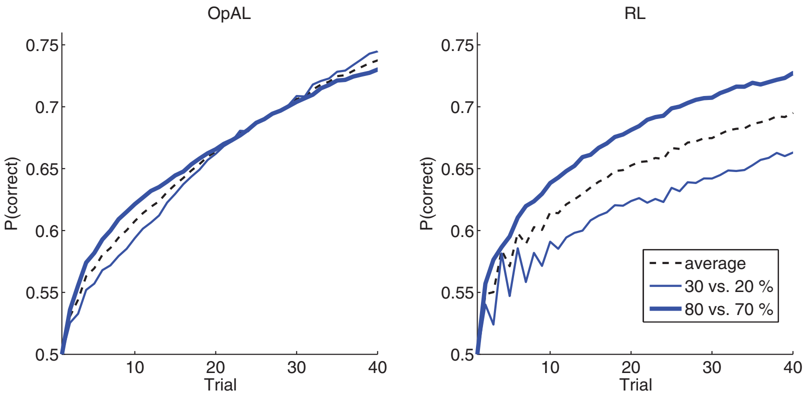

Figure 13. Discrimation learning. Simulation of opponent actor learning (OpAL) and reinforcement learning (RL) model on a discrimination learning task, with optimized parameters, averaged over 10,000 iterations. One pair of options lead to reward 80% versus 70% of the time, the other 30% versus 20%. Graphs show average probability of selecting the "correct" option, namely, the 80% or 30% options. See the online article for the color version of this figure.

resentation of the values (see supplemental results Appendix). Moreover, straightforward approaches to dynamically changing the β parameter in simple RL as a function of time did not accomplish the same goal. Thus, this analysis shows that having a dual system with nonlinear update rules in OpAL affords better performance in discrimination learning, by virtue of including a mechanism for detecting the appropriate actor weights that contribute to choice.

Empirically, although we have made the case that the neurobiology is more in accordance with OpAL than standard RL models, this normative analysis raises the question of whether human behavior accords with the predictions of OpAL. In contrast to OpAL, standard RL models predict a clear difference in accuracy in lean versus rich stimuli. Although we are not aware of experiments testing the precise design we simulated above, a related rich versus lean design was employed by Pessiglione et al. (2006), showing no performance difference between conditions. Nevertheless, when simulating this experiment with standard RL models and best fit parameters provided in Pessiglione et al. (2006), we obtained a strong asymmetry in performance, again with better performance for the rich (more rewarded) pair, contrary to that observed empirically. In contrast, with appropriate parameters OpAL reproduced the observed pattern of data across both conditions, for the same reasons as described baove. Thus, while limited, available data indicates that RL models are inconsistent with learning curves observed in human subjects and that OpAL overcomes this issue.

Discussion

We have presented a new computational model to simultaneously account for learning and incentive/performance effects of striatal dopamine. Existing models have largely focused on one aspect or the other, but we show here how the combination of these features and their interactions are needed to explain a wide range of data. This model is based on an actor-critic architecture with a few key novel features. First, it relies on opponent actor weights that independently track the attractiveness and aversiveness of options, with a choice policy depending on a weighted competition between those signals. Second, these actor weights mimic striatal plasticity rules by including a multiplicative Hebbian term, which leads to a nonlinear representation of values with separate emphasis of attractive or aversive parts of the scale.

These features combined together provide the flexibility needed to account for a wide array of data. Indeed, due to the nonlinear learning rule, the *G* and *N* weights come to differentiate distinct aspects of value representations, which are stored separately, and dynamically recombined with parameterized weighting as a function of dopamine levels during choice. This system thus allows a great flexibility in the way past learned information is expressed as a choice in different situations, as a function of motivational state. Indeed, we showed that OpAL could account for learned choice biases and for how dopaminergic manipulation, either during learning, or at the time of choice, could shift those biases. We also showed that OpAL could account for complex interactive effects of performance and learning, capturing empirical findings in the rotarod motor skill learning task but also leading to novel predictions for reward-based decision making: Choice incentive effects were stronger when choice is made in the same dopaminergic state as that during learning. Similarly, the model predicts that both learning and incentive effects can impact effort-based decision making, that these effects amplify each other, and that the relative impact on one or the other process is dependent on the overall degree of effort required (Figure 9).

The Multiple Computational Roles of Dopamine in Learning and Choice

Our new OpAL model relies critically on well-established neurological data: the existence of two opposing, seemingly redundant systems: the D1-direct and D2-indirect pathway (modeled here through *G* and *N* weights) and the role of dopamine in reinforcement learning but also in choice and motivation. It has been a matter of speculation to figure out why this double encoding existed in the brain and what computational advantage it offered. It is also unclear why dopamine seems to play so many roles, functionally encoding signals such as prediction error, motivational salience, etc. Our model provides no direct answers to these questions, but opens the way to investigating them further. By allowing flexible exploration of behavior in different environments as well as normative analysis, it offers some clues into the potential roles of these key characteristics.

In particular, our simulations investigated the interaction between the role of dopamine for learning and for choice. In the probabilistic selection task, we showed that overall performance became strongly impaired if dopaminergic status was different between learning and testing. This may provide a clue as to why the same neurotransmitter is apparently used to modulate incentive and learning: if any factor (disease, development, etc.) causes dopamine levels to be either high or low, that will lead to an asymmetry in learning in the *G* versus *N* weights but will also allow the choice function to rely preferentially on the weights that have accumulated the most useful information. A coupling between the control of the balance between the opponent systems during learning and during testing might thus optimize choice.

Our model also proposes a new understanding for the seemingly redundant coding in direct and indirect pathways of choice values. Indeed, our simulations show that it allows better performance in a discrimination learning task, when having to learn away from an optimistic initial estimation brings on a conflict between exploring and exploiting different choices. Thus, the presence of an opponent system enforces a temporarily suboptimal representation of preferences, that nevertheless allows for better long term estimation: It obviates a tradeoff between exploration and exploitation for environments with sparse reward schedules. We have also shown that the opponent systems represented redundant signals but that their representations were more precise in different domains of value: *G*-weights emphasize differences between *good* options, while *N* weights emphasize differences between *bad* options. Speculatively, having the flexibility of setting the emphasis on either of those signals might allow the system to dynamically decide, based on a motivational state that could be encoded by dopamine levels at the time of choice (β weights), which information to put more stock on to decide on a policy: Do costs matter more or gains? This is a potential benefit of the opponent structure that will need to be explored in further research.

Relationship to Neural Network Models

OpAL was built to mimic some of the core principles embedded within more complex dynamical neural network models of the corticostriatal circuitry in reinforcement learning and reward-based decision-making articles (e.g., Beeler et al., 2012; Collins & Frank, 2013; Frank, 2005; Wiecki et al., 2009), by simplifying these neurobiologically inspired neural network models into an algorithmic form. This raises the question of what advances it provides in comparison to this class of models: Indeed, as such, it qualitatively accounts for a similar array of results simulated with this neural network model. Nevertheless, we believe our work offers several new contributions. Although OpAL does not consider neural dynamics among multiple basal ganglia nuclei, thalamus, and cortex, it presents several advantages over the neural network version, which we detail now.

First, prior articles showing neural network simulations of some of the tasks we modeled here (e.g., the probabilistic selection task) emphasized the role of dopamine on learning. Although some articles alluded to performance effects on choice incentive by referring to the effects of dopamine on Go versus NoGo activity in the striatum of that model, the differentiable roles of this choice incentive effect has not been formally investigated in a publication, except in the specific case of motor skill learning and performance (Beeler et al., 2012), as in the rotarod simulations we include here. But the role of dopamine in modulating cost/benefit choice incentives in reward based tasks, including reinforcement learning and effort-based decision making tasks, has never been reported or accounted for by ours or any other existing model. Furthermore, previous simulations of dopamine manipulations in neural network studies have generally used one set of parameters simulating an increase or decrease of DA by fixed values. Conversely, OpAL allows full characterization of effects across the entire range of parameters for both learning and motivation and how they interact.

On a pragmatic side, as a simpler, low parameter algorithmic model, OpAL provides strong advantages for analysis. It can be quantitatively fit to empirical data, and affords simple theoretical analysis: Understanding the information represented by model variables is straightforward, the dependence of the model's dynamics on model parameters and task environments can be easily and exhaustively analyzed, and we can identify regimes that optimize its performance in various environments. Thus, while the neural network version focuses on more detailed mechanisms, its high level functions are not as transparent. The current model thus provides a better understanding of how various components of the system interact with each other, leading to novel predictions described in this report, which had not previously been elucidated. For example, we assess the influence of these modulations not only on one particular version of a task (e.g., the probabilistic selection task with fixed reward probabilities) but to a generalized version, showing novel predictions about how the standard results should change as a function of task probabilities (Figure 8). Similarly, the model predicts differential effects of dopamine manipulation depending on the experimental stage of the manipulation: Whereas effort tasks typically manipulate dopamine after learning about effort has already occurred, our model predicts that manipulation during the learning phase would interact with those of manipulation during the performance phase and, moreover, that their differential impact would depend on the level of effort required (Figure 9).

Finally, because OpAL is indeed simpler (far fewer parameters) than the neural network model, it is amenable to quantitative fitting provided experimental designs that are rich enough to distinguish between learning and motivational effects. Moreover, this simpler model affords the type of normative analysis we provide in the article.

Relationship to Existing Models

To our knowledge, other computational models cannot account for the range of data presented here. In particular, classic reinforcement learning models track true expected value and make choices based on relative differences in expected value, and as such cannot reproduce any choice biases observed experimentally.

Previous attempts at modeling asymmetries in learning used a classic RL model without separating *G* and *N* systems but instead simply assumed different learning rates for positive and negative prediction errors within a single value system (e.g., Doll et al., 2011; Frank, Moustafa, et al., 2007). That model was able to account for some asymmetries due to learning, but the effects were rather counterintuitive: Better Choose-A performance was associated with *lower* learning rates from positive prediction errors, and better Avoid-B was associated with lower learning rates from negative prediction errors. Moreover, systematic investigation of that model reveals that while these effects hold in general, they are quite nonlinear (see Appendix, Figure A3C). More important, allowing only for asymmetry in learning rates precludes the possibility of differentially expressing preferences during performance: Since this method still integrates all information into a single value, it fails to provide the flexibility needed to account for potential incentive performance effects (see Appendix for detailed results). Further, the current model more closely aligns with the biology and neural models of differential *G* and *N* systems that operate during learning and choice.

A model by McClure et al. (2003) included in a single framework the reinforcement learning and incentive theories of dopamine by assuming that the phasic dopamine signal at stimulus presentation, encoding choice incentive, corresponded to future expected value and modulated the gain in the softmax choice function. While that model accounted for some incentive effects in the appetitive domain, it predicts systematically lower performance in choosing between less valuable stimuli (which would amount to low gain). This is the opposite of what is observed in the low dopaminergic state empirically and in OpAL.

Computational models that focus on the motivation or incentive part of the dopamine literature typically focus on effort or response vigor for a single action (Dayan, 2012; Niv et al., 2007) and, as such, have not addressed the potential influence of dopamine on relative emphasis on prospective gains or losses. Similarly, models that include Pavlovian to instrumental transfer (PIT) effects (e.g., Huys et al., 2011) account for the impact of stimulus values on the overall invigoration of action but do not predict differential effects on choice of positively versus negative valenced options. Berridge and Zhang's model (Zhang et al., 2009) tries to account for all learning and incentive effects through a single incentive mechanism, reflected by a κ parameter. Although this model does account for a wide range of data, the κ model also does not address the tendency for low DA levels to enhance performance in avoiding (or differentiating between) negatively valenced options. Their model does simulate the possibility for an animal to "want" what is not expected to be liked, nor remembered to be "liked" (i.e., a valence reversal), by differentially impacting motivational state at the time of choice to inflate the effective value of an option that had been encoded as mostly aversive. Our model can also account for these same effects, where an asymmetry in learning can be reversed by an asymmetry in choice incentive (Figure A4). Moreover the κ model required an alteration to use a log-based transformation that only applies in the case of a valence reversal, whereas the OpAL model captures this naturally without alteration.

Limitations

By simplifying the neurobiologically inspired neural network model from Frank (2005) into an algorithmic form, we have introduced some limitations. One key limitation is that we have focused fully on the actor (or dorsal striatal) part of the problem, and neglected the critic side (ventral striatum). Dopamine should also have effects on the critic, and a more detailed model, beyond the scope of this work, should incorporate them. However, we are confident that modification of this aspect would not significantly alter the qualitative results presented here. Indeed, we simulated different versions of the critic part of our model (for example including asymmetric gain/loss learning rate) and obtained qualitatively similar results, if quantitatively different.

We also simplified some aspects in the actor part of the model. For example, it could be reasonably expected that plasticity constants are not equal for potentiation or depression in each pathway. This could be modeled by including separate gain and a loss learning rates in each system G and N, thus leading to four actor learning rates. While this could potentially include more flexibility in the way biases express themselves, we chose to analyze a simpler version, because we could not find experimental data that would allow us to successfully dissociate roles of each of these learning rates. We also do not consider the roles of the subthalamic nucleus and modulations of decision thresholds that remain an important aspect of the neural circuit (Frank & O'Reilly, 2006; Wiecki & Frank, 2013).

Although our model can account for effects of dopamine on different kinds of effort tasks, including lever pressing tasks and the T-maze with barrier, our model cannot at this point capture some finer grained notions of effort. For example, we do not model the strength or force with which a response is emitted—although intuitively, the relative G to N difference for each action under consideration should determine not only which action wins and its response time but also the degree of boost provided to motor thalamocortical populations, which may affect its "strength." Future work will examine this notion. Other potential directions for extensions and elaborations includes accounting more precisely for timing in effort paradigms (in particular in the fixed ratio lever task, using probabilities as a proxy for fixed ratio prevents analysis of dynamic changes to vigor in expectation of reward) or for other indicators of effort.

Contrary to other computational theories such as (Dayan, 2012; Niv et al., 2007), we do not derive the normative optimization for dopamine levels given some objective function. Instead, we explore the effects of differing dopamine levels on the opponent actor weights via modulation of parameters β_G and β_N. While this allows us to capture effects of dopamine manipulations on choice preferences that are not accounted for by other model is, we have not yet explored how these levels should change as a function of task context so as to optimize performance. For example, there may be conditions under which it is advantageous to emphasize G weights over N weights and vice versa. Future work should identify these conditions and whether empirically, dopamine levels are adjusted accordingly.

Conclusion

The OpAL model provides a new biologically grounded reinforcement learning framework that accounts for a wide array of data linking behavior in learning and performance tasks to contributions of dopaminergic and striatal direct and indirect pathway neurons. It makes further testable predictions and provides some clues to the understanding of important open questions, such as why we have redundant pathways in the basal ganglia and why dopaminergic function is so omnipresent. Future research will investigate predictions and use the OpAL framework to expand our understanding of this complex system.

References

Aberman, J. E., & Salamone, J. D. (1999). Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience, 92,* 545–552. doi:10.1016/S0306-4522(99)00004-4

Amtage, J., & Schmidt, W. J. (2003). Context-dependent catalepsy intensification is due to classical conditioning and sensitization. *Behavioural Pharmacology, 14,* 563–567. doi:10.1097/00008877-200311000-00009

Arias-Carrión, O., Stamelou, M., Murillo-Rodríguez, E., Menéndez-González, M., & Pöppel, E. (2010). Dopaminergic reward system: A short integrative review. *International Archives of Medicine, 3,* 24.

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47,* 129–141. doi:10.1016/j.neuron.2005.05.020

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

Beeler, J., Frank, M., McDaid, J., & Alexander, E. (2012). A role for dopamine-mediated learning in the pathophysiology and treatment of Parkinson's disease. *Cell Reports, 2,* 1747–1761.

Beeler, J. A., Daw, N., Frazier, C. R. M., & Zhuang, X. (2010). Tonic dopamine modulates exploitation of reward learning. *Frontiers in Behavioral Neuroscience*. doi:10.3389/fnbeh.2010.00170

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology, 191,* 391–431. doi:10.1007/s00213-006-0578-x

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

Bódi, N., Kéri, S., Nagy, H., Moustafa, A., Myers, C. E., Daw, N., . . . Gluck, M. A. (2009). Reward-learning and the novelty-seeking personality: A between- and within-subjects study of the effects of dopamine agonists on young Parkinson's patients. *Brain: A Journal of Neurology, 132,* 2385–2395. doi:10.1093/brain/awp094

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review, 120,* 190–229. doi:10.1037/a0030852

Cools, R., Frank, M. J., Gibbs, S. E., Miyakawa, A., Jagust, W., & D'Esposito, M. (2009). Striatal dopamine predicts outcome-specific reversal learning and its sensitivity to dopaminergic drug administration. *The Journal of Neuroscience, 29,* 1538–1543. doi:10.1523/JNEUROSCI.4467-08.2009

Cousins, M., Atherton, A., Turner, L., & Salamone, J. (1996). Nucleus accumbens dopamine depletions alter relative response allocation in a T-maze cost/benefit task. *Behavioural Brain Research, 74,* 189–197.

Cousins, M. S., & Salamone, J. D. (1994). Nucleus accumbens dopamine depletions in rats affect relative response allocation in a novel cost/benefit procedure. *Pharmacology, Biochemistry and Behavior, 49,* 85–91. doi:10.1016/0091-3057(94)90460-X

Dayan, P. (2012). Instrumental vigour in punishment and reward. *European Journal of Neuroscience, 35,* 1152–1168. doi:10.1111/j.1460-9568.2012.08026.x

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective & Behavioral Neuroscience, 8,* 429–453. doi:10.3758/CABN.8.4.429

Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience, 31,* 6188–6198.

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research, 1299,* 74–94. doi:10.1016/j.brainres.2009.07.007

Farrar, A. M., Font, L., Pereira, M., Mingote, S., Bunce, J. G., Chrobak, J. J., & Salamone, J. D. (2008). Forebrain circuitry involved in effort-related choice: Injections of the GABAA agonist muscimol into ventral pallidum alter response allocation in food-seeking behavior. *Neuroscience, 152,* 321–330. doi:10.1016/j.neuroscience.2007.12.034

Farrar, A. M., Segovia, K. N., Randall, P. A., Nunes, E. J., Collins, L. E., Stopper, C. M., . . . Salamone, J. D. (2010). Nucleus accumbens and effort-related functions: Behavioral and neural markers of the interactions between adenosine A2A and dopamine D2 receptors. *Neuroscience, 166,* 1056–1067. doi:10.1016/j.neuroscience.2009.12.056

Floresco, S. B., Tse, M. T. L., & Ghods-Sharifi, S. (2008). Dopaminergic and glutamatergic regulation of effort- and delay-based decision making. *Neuropsychopharmacology, 33,* 1966–1979.

Frank, M. J. (2005). Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *Journal of Cognitive Neuroscience, 17,* 51–72. doi:10.1162/0898929052880093

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 16311–16316. doi:10.1073/pnas.0706111104

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

Frank, M. J., Santamaria, A., Reilly, R. C. O., & Willcutt, E. (2007). Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder. *Neuropsychopharmacology, 32,* 1583–1599.

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

Gerfen, C. R. (2000). Molecular effects of dopamine on striatal-projection pathways. *Trends in Neurosciences, 23* (Suppl), S64–S70. doi:10.1016/S1471-1931(00)00019-7

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2010). Neural mechanisms of acquired phasic dopamine responses in learning. *Neuroscience and Biobehavioral Reviews, 34,* 701–720. doi:10.1016/j.neubiorev.2009.11.019

Hikida, T., Kimura, K., Wada, N., Funabiki, K., & Nakanishi, S. (2010). Distinct roles of synaptic transmission in direct and indirect striatal pathways to reward and aversive behavior. *Neuron, 66,* 896–907. doi:10.1016/j.neuron.2010.05.011

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Computational Biology, 7,* e1002028. doi:10.1371/journal.pcbi.1002028

Jocham, G., Klein, T. A., & Ullsperger, M. (2011). Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. *The Journal of Neuroscience, 31,* 1606–1613.

Klein, A., & Schmidt, W. J. (2003). Catalepsy intensifies context-dependently irrespective of whether it is induced by intermittent or chronic dopamine deficiency. *Behavioural Pharmacology, 14,* 49–53.

Kravitz, A. V., Freeze, B. S., Parker, P. R. L., Kay, K., Thwin, M. T., Deisseroth, K., & Kreitzer, A. C. (2010). Regulation of Parkinsonian motor behaviours by optogenetic control of basal ganglia circuitry. *Nature, 466,* 622–626. doi:10.1038/nature09159

Kravitz, A. V., Tye, L. D., & Kreitzer, A. C. (2012). Distinct roles for direct and indirect pathway striatal neurons in reinforcement. *Nature Neuroscience, 15,* 816–818. doi:10.1038/nn.3100

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience, 14,* 154–162.

McClure, S. M., Daw, N. D., & Read Montague, P. (2003). A computational substrate for incentive salience. *Trends in Neurosciences, 26,* 423–428. doi:10.1016/S0166-2236(03)00177-2

Mingote, S., Font, L., Farrar, A. M., Vontell, R., Worden, L. T., Stopper, C. M., . . . Salamone, J. D. (2008). Nucleus accumbens adenosine A2A receptors regulate exertion of effort by acting on the ventral striatopallidal pathway. *The Journal of Neuroscience, 28,* 9037–9046.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience, 16,* 1936–1947.

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

Moustafa, A. A., Sherman, S. J., & Frank, M. J. (2008). A dopaminergic basis for working memory, learning and attentional shifting in Parkinsonism. *Neuropsychologia, 46,* 3144–3156. doi:10.1016/j.neuropsychologia.2008.07.011

Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y., & Hikosaka, O. (2004). Dopamine neurons can represent context-dependent prediction error. *Neuron, 41,* 269–280. doi:10.1016/S0896-6273(03)00869-9

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

Nomoto, K., Schultz, W., Watanabe, T., & Sakagami, M. (2010). Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *The Journal of Neuroscience, 30,* 10692–10702.

Nunes, E. J., Randall, P. A., Santerre, J. L., Given, A. B., Sager, T. N., Correa, M., & Salamone, J. D. (2010). Differential effects of selective adenosine antagonists on the effort-related impairments induced by dopamine D1 and D2 antagonism. *Neuroscience, 170,* 268–280. doi:10.1016/j.neuroscience.2010.05.068

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

Ott, D. V. M., Ullsperger, M., Jocham, G., Neumann, J., & Klein, T. A. (2011). Continuous theta-burst stimulation (cTBS) over the lateral prefrontal cortex alters reinforcement learning bias. *NeuroImage, 57,* 617–623. doi:10.1016/j.neuroimage.2011.04.038

Palminteri, S., Boraud, T., Lafargue, G., Dubois, B., & Pessiglione, M. (2009). Brain hemispheres selectively track the expected value of contralateral options. *The Journal of Neuroscience, 29,* 13465–13472.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature, 442,* 1042–1045. doi:10.1038/nature05051

Ratcliff, R., & Frank, M. J. (2011). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation, 24,* 1186–1229. doi:10.1162/NECO_a_00270

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

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience, 10,* 1615–1624. doi:10.1038/nn2013

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

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

Satoh, T., Nakai, S., Sato, T., & Kimura, M. (2003). Correlated coding of motivation and outcome of decision by dopamine neurons. *The Journal of Neuroscience, 23,* 9913–9923.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience, 27,* 12860–12867.

Schultz, W. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599. doi:10.1126/science.275.5306.1593

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science, 321,* 848–851. doi:10.1126/science.1160575

Shiner, T., Seymour, B., Wunderlich, K., Hill, C., Bhatia, K. P., Dayan, P., & Dolan, R. J. (2012). Dopamine and performance in a reinforcement learning task: Evidence from Parkinson's disease. *Brain: A Journal of Neurology, 135,* 1871–1883. doi:10.1093/brain/aws083

Smith, K. S., Berridge, K. C., & Aldridge, J. W. (2011). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *Proceedings of the National Academy of Sciences of the United States of America, 108,* E255–E264. doi:10.1073/pnas.1101920108

Smittenaar, P., Chase, H. W., Aarts, E., Nusselein, B., Bloem, B. R., & Cools, R. (2012). Decomposing effects of dopaminergic medication in Parkinson's disease on probabilistic action selection: Learning or performance? *European Journal of Neuroscience, 35,* 1144–1151. doi:10.1111/j.1460-9568.2012.08043.x

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences, 30,* 228–235. doi:10.1016/j.tins.2007.03.008

Sutton, R., & Barto, A. (1998). Reinforcement learning. *Journal of Cognitive Neuroscience, 11,* 126–134. doi:10.1162/089892999563184

Tai, L.-H., Lee, A. M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience, 15,* 1281–1289. doi:10.1038/nn.3188

Treadway, M. T., Buckholtz, J. W., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., . . . Zald, D. H. (2012). Dopaminergic mechanisms of individual differences in human effort-based decision-making. *The Journal of Neuroscience, 32,* 6170–6176.

Wassum, K. M., Ostlund, S. B., Balleine, B. W., & Maidment, N. T. (2011). Differential dependence of Pavlovian incentive motivation and instrumental incentive learning processes on dopamine signaling. *Learning & Memory, 18,* 475–483. doi:10.1101/lm.2229311

Wiecki, T. V., & Frank, M. J. (2013). A computational model of inhibitory control in frontal cortex and basal ganglia. *Psychological Review, 120,* 329–355. doi:10.1037/a0031542

Wiecki, T. V., Riedinger, K., von Ameln-Mayerhofer, A., Schmidt, W. J., & Frank, M. J. (2009). A neurocomputational account of catalepsy sensitization induced by D2 receptor blockade in rats: Context dependency, extinction, and renewal. *Psychopharmacology, 204,* 265–277. doi:10.1007/s00213-008-1457-4

Zhang, J., Berridge, K. C., Tindell, A. J., Smith, K. S., & Aldridge, J. W. (2009). A neural computational model of incentive salience. *PLoS Computational Biology, 5,* e1000437. doi:10.1371/journal.pcbi.1000437

(*Appendix follows*)

Appendix

Supplemental Methods: Normative Analysis

Task

Models were simulated over 50 trials per pairs of options with reinforcement $r = 1$ or $r = 0$—good options: $p(r = 1) = 0.8$ and 0.7, bad options: $p(r = 1) = 0.3$ and 0.3. We define performance as the probability of choosing the objective good option (0.8 and 0.3).

To optimize the model parameters, we ran an optimization (matlab *fminsearch* with 20 randomly chosen starting points) procedure on the average performance of 1,000 simulations, keeping the same 1,000 random seeds. Those seeds were also kept across different model optimizations, to ensure that differences were not due to chance in the stochastic policies.

We also report optimization of a classic *delta rule* RL model, and three parameter OpAL model. The RL model estimated the value of each of the four options according to the Equation 1. These values were used for choice in a softmax with probability of choosing s_i from the pair (s_1, s_2) equal to

$$p(s_i) = \frac{exp(\beta V_{s_i})}{exp(\beta V_{s_1}) + exp(\beta V_{s_2})}.$$

V was initialized at chance level of reward, 0.5. Optimized parameters were $\alpha = 0.24$ and $\beta = 27.4$.

OpAL model was optimized with three parameters (α_C, $\alpha_G = \alpha_N$, $\beta_G = \beta_N$). Critic and actors are initialized at chance ($V = 0.5$, $G = N = 1$). Optimized parameters were $\alpha_C = 0.035$, $\alpha_{GN} = 0.98$ and $\beta_{GN} = 1.5$. Although RL model was optimized with less parameters, this could not account for the difference observed: More flexible RL models (including asymmetric gain/loss learning rates) were also optimized an did not match OpAL performance. A model accumulating frequencies of reward for each option also performed worse than OpAL.

Alternative Models

Win-loss model. A frequently used model to account for asymmetrical learning effects of dopamine is a win-loss reinforcement learning model. In this model, a single value estimate is updated according to a classic learning rule, but with different learning rates depending on the sign of the prediction error:

$$\text{if } \delta(t) > 0 \qquad V(t+1) = V(t) + \alpha_W \delta(t) \tag{A1}$$

$$\text{if } \delta(t) \le 0 \qquad V(t+1) = V(t) + \alpha_L \delta(t) \tag{A2}$$

where $\delta(t) = r(t) - V(t)$ is the prediction error, and α_W and α_L are gain and loss learning rates, respectively.

In previous quantitative fitting studies in humans, individual differences in fitted α_W and α_L were related to Choose-A and Avoid-B performance, respectively, Frank, Moustafa, et al. (2007). Simulations in Figure A1 confirm that with different values of α_W and α_L, the representation of option values is distorted: With higher α_W, estimates are lifted. However, note that this implies a loss of sensitivity in good options, while increasing α_L stretches their representation and thus increases sensitivity to learning from positive rewards. This explains that the bias for the probabilistic learning task, obtained from asymmetry in learning parameters, is in a counterintuitive direction: better Choose-A performance is associated with relative *decrease* in α_W, whereas better Avoid-B performance is associated with decrease in α_L (see Figures A1C and A1D), as found empirically in Frank, Moustafa, et al. (2007). Moreover, the relationship is nonlinear. In contrast, OpAL provides more straightforward interpretation, where α_G is positively (and monotonically) related to Choose-A performance, and similarly for α_N and Avoid-B. Moreover, as emphasized in the text, the win-loss model provides no mechanisms to modulate choice incentive at the time of decision.

OpAL without Hebbian update rule. To demonstrate the necessity of the modulation by actor G and N of the learning rules, we also ran simulations without this modulation. In that model, all equations are similar to OpAL, without the multiplicative term:

$$G_a(t+1) = G_a(t) + [\alpha_G] \times \delta(t) \tag{A3}$$

$$N_a(t+1) = N_a(t) + [\alpha_N] \times [-\delta(t)] \tag{A4}$$

Simulations are shown in supplemental Figure A2. In this model, G, N, and *Act* are linear functions of the critic value (see

(*Appendix continues*)

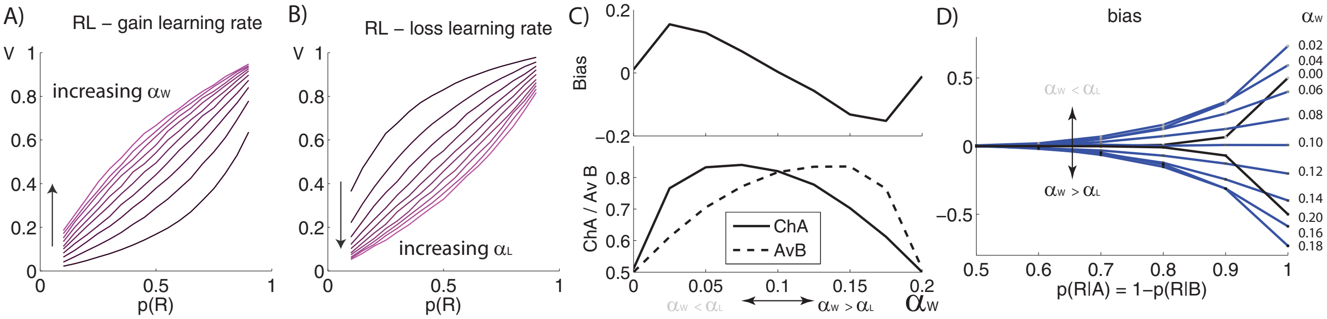

Figure A1. Gain-loss learning rate reinforcement learning (RL) model. All values are final values after 100 trials, averaged over 1,000 simulations. A, B. Value estimates (V) for various probabilities of reward $p(R)$. Lighter shade and arrow direction indicate higher parameter values (A $\alpha_L = 0.1$, $\alpha_W \in [0.02, 0.2]$; B $\alpha_G = 0.1$, $\alpha_L \in [0.02, 0.2]$). C, D. Probabilistic selection task (see main text Figure 8 for comparison), with α_W, $\alpha_L \in [0, 0.2]$ and $\alpha_W + \alpha_L = 0.2$. C. Simplified task with $p(R \mid A) = 0.8$, $p(R \mid B) = 0.2$. D. Darker dots indicate bias toward W versus L, while lighter dots indicate bias toward L versus W. Black curves are highlighted for either α_G or $\alpha_L = 0$. Note that bias toward avoiding increases with increased α_W. ChA indicates Choose-A performance, and AvB indicates Avoid-B performance. See the online article for the color version of this figure.

Figures A2A and A2B, compared to OpAL in Figure 3). Thus, any information about differences in positive outcomes is contained symmetrically by large *G* values and small *N* values, and vice versa for negative outcomes. This prevents the expression of any bias in the probabilistic selection task, with either learning rate or β asymmetry (see Figure A2C vs. main text Figure 3). Although similar effort effects can be obtained to the normal OpAL (Figure A2D, compared to main text Figure 9 left), by simply modulating the overall contribution of *G* to *N* weights for a single action, the model fails to show the triple interaction described in the main text, which relies on the

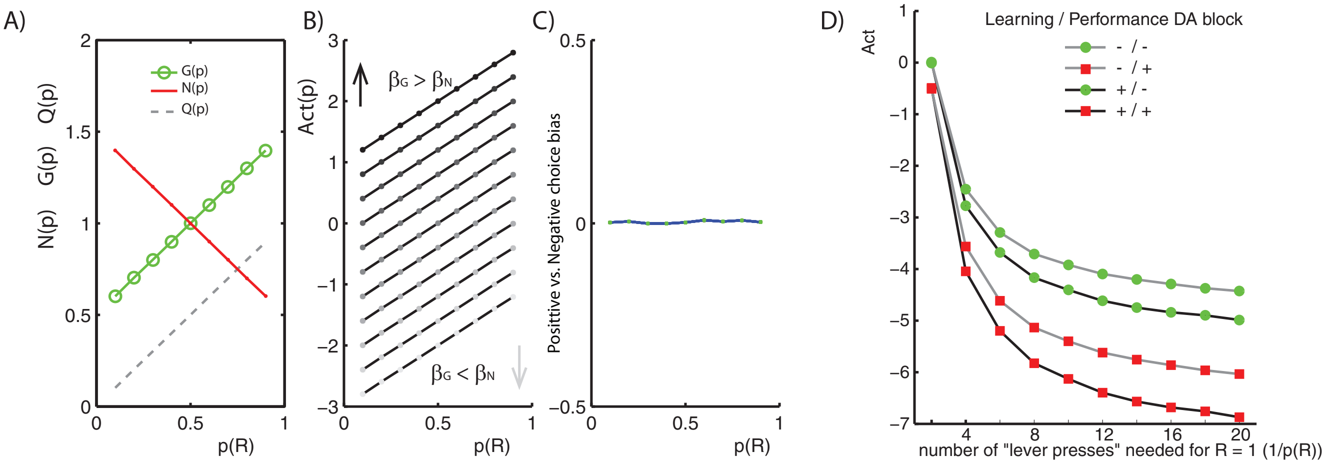

Figure A2. No-Hebb opponent actor learning (OpAL) model. A–C. Model values for different probabilities of reward p(R) (cf. main text Figure 3): critic value *Q*, *G* and *N* weights, and actor values Act. Note in particular that even when changing β_G versus β_N or α_G versus α_N, no bias in selection is observed (C). D. Effort task, with same parameters as normal OpAL (see main text Figure 3). No interaction effect is observed. DA = dopamine. See the online article for the color version of this figure.

(*Appendix continues*)

Hebbian multiplicative term. Similarly, without the Hebbian term, the model fails to capture the full pattern of data in the rotarod motor skill learning task (not shown). In particular, it does not capture the critical signature of slowed recovery after dopamine blockade observed in Beeler et al. (2012) and prior studies.

Critic Dependent Effects

Critic learning rate effects. In addition to observing the effects of actor learning rates, we also examined the specific contributions of the critic. First, we parametrically varied the critic initialization value. Simulations (Figure A3 top) showed that pessimistic critic initialization ($V(0) < 0.5$) lead to stronger G than N actor weights, reflecting the accumulation of more positive prediction errors than negative ones. In particular, this leads to higher actor weights and a bias toward Choose-A over Avoid-B (blue line). The opposite happens when the critic is initialized too optimistically ($V(0) > 0.5$): In this case, an overbalance of negative prediction errors favors N weights over G.

In a separate set of simulations (see Figure A3 bottom), we parametrically manipulated the critic learning rate α_C. Results showed that, although critic learning rates do not modify the asymptotic critic value, they do affect asymptotic actor weights. Specifically, smaller critic learning rates led to emphasis in the modulation seen in normal dynamics, such that *good* stimuli are perceived with even stronger G weights, "bad" stimuli with even smaller G weights, and vice versa for N weights. This is because

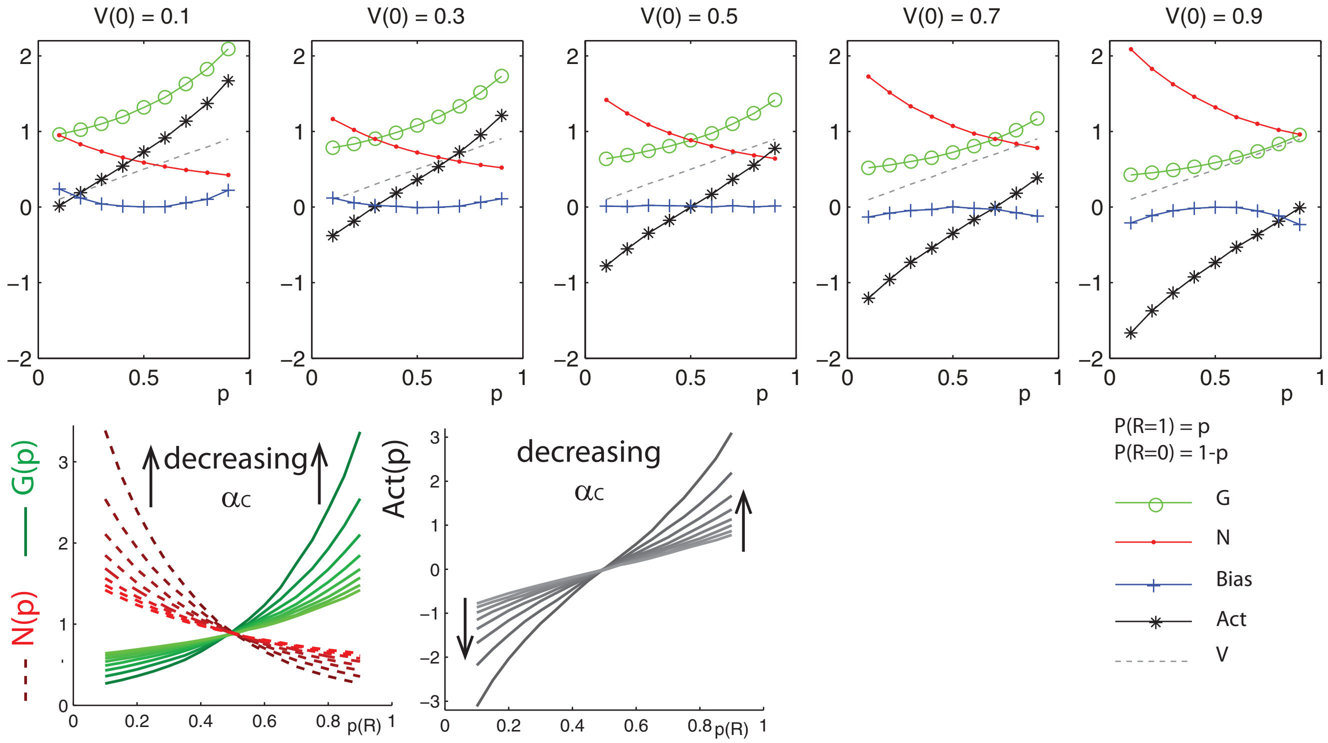

Figure A3. Critic effects on opponent actor learning (OpAL). All values are final values after 100 trials, averaged over 1,000 simulations. Top graphs: Dashed gray line is critic value V, which always coincides with expected value. The open circles line (green line online) is G weight, closed circles line (red online) is N weight, and black stars line is actor value $Act = G - N$. The crosses line (blue online) indicates the bias that would be observed compared to an option presenting reward probability $1 - p(R)$. Bottom graphs: Critic learning rate α_C. Decreasing α_C induces slower convergence of critic, leading to longer accumulation of prediction error-related biases in actor weight, thus exaggerating all biases and nonlinearities. See the online article for the color version of this figure.

(*Appendix continues*)

with slower critic convergence, prediction errors accumulate for a longer time period in the actor, which thus accumulates stronger biases.

OpAL stability. The critic equation can be written as

$$V(t+1) = V(t) + \alpha_C \delta(t) \tag{A5}$$

$$= V(0) + \sum_{i=1}^{t} \alpha_C \delta(t) \tag{A6}$$

In stationary environments with sufficient exploration, this estimate converges in probability (under some reasonable assumptions) to the true expected reward. In particular, this implies that the expected value of the prediction error $\delta(t)$ converges to 0.

The actor weight equations can also be written as a function of the series of prediction errors:

$$G(t+1) = G(t) + \alpha_G \delta(t) G(t) \tag{A7}$$

$$= G(t) \times (1 + \alpha_G \delta(t)) \tag{A8}$$

$$= G(0) \times \prod_{i=1}^{t} (1 + \alpha_G \delta(i)) \tag{A9}$$

Similarly, we obtain $N(t+1) = N(0) \times \prod_{i=1}^{t} (1 - \alpha_N \delta(i))$.

In stationary environments, under the same assumptions that ensure convergence of classic RL models, the actor weights cannot diverge. Indeed, we can show that

$$log(G(t+1)) = \log(G(0)) + \sum_{i=0}^{t} \log(1 + \alpha_G \delta(t)) \tag{A10}$$

$$< \; log(G(0)) + \alpha_G \sum_{i=1}^{t} \delta(t) \tag{A11}$$

$$= \; log(G(0)) + \frac{\alpha_G}{\alpha_C}(V(t+1) - V(0)) \tag{A12}$$

Thus, G is upper bound by a function of the critic and lower bound by 0 as a constraint of the model, representing the notion that firing rates cannot go below zero. The same derivation can be applied for N.

Optimization of Discrimination Learning

RL. Here, we show mathematically why an RL model learns slower for the 20–30 than for the 70–80 case. Let A and B denote the two options of a pair. At time point t, the model chooses option X = A or B with probability from a softmax policy π_X and receives reward 1 with probability p_X. We are interested in how the difference in estimated values between the options, $\Delta Q = Q_A - Q_B$, which determines performance, changes from the outcome of the trials. This can be seen in Equation A13. Thus, the expected value of $\Delta Q(t+1)$, given all values at time t is

$$E(\Delta Q(t+1) \mid \text{values at t}) = \pi_B[(1-\alpha)Q_B - Q_A + \alpha p_B] + \pi_A[Q_B - (1-\alpha)Q_A - \alpha p_A]$$

Taking into account that $\pi_A = 1 - \pi_B$, then writing $Q_B = Q_A + \Delta Q(t)$, and $p_B = p_A + \epsilon$, we come to simplify this as

$$E(\Delta Q(t+1)) = \alpha \pi_B \epsilon + \alpha(2\pi_B - 1)(p_A - Q_A) + \Delta Q(t)(1 - \alpha \pi_B)$$

Note that only the middle term of this sum depends on the absolute value of the options (in addition to their relative values), and as such it is the term of most interest to explain the difference in learning between the 20–30 case and the 70–80 case. Let's assume that B is the better option in the pair, with $\epsilon > 0$. In the *lean* case $p_A = 0.2$ and should initially be smaller than the estimated value Q_A, which is initialized at 0.5. Thus, if $\pi_B > 0.5$ (reflecting correct learning that B is the better option), the middle factor is negative and thus slows the expected increase in discrimination between values. In contrary in the *rich* case, $p_A = 0.7$, $\pi_B > 0.5$ makes the discrimination increase faster, because initially $p_A - Q_A > 0$. Thus, to a first approximation, this shows why learning is slower in the 20–30 case than in the 70–80 case. Of course, one can initialize values at other points than 0.5, but there is no unique value that will solve this problem for any arbitrary probabilistic discrimination.

$$\begin{aligned}
\Delta Q(t+1) &= (1-\alpha)Q_B + \alpha - Q_A && \text{with probability} \quad \pi_B \times p_B \\
&= (1-\alpha)Q_B - Q_A && \pi_B \times (1 - p_B) \\
&= Q_B + \alpha - [(1-\alpha)Q_A + \alpha] && \pi_A \times p_A \\
&= Q_B + \alpha - [(1-\alpha)Q_A] && \pi_A \times (1 - p_A)
\end{aligned} \tag{A13}$$

(*Appendix continues*)

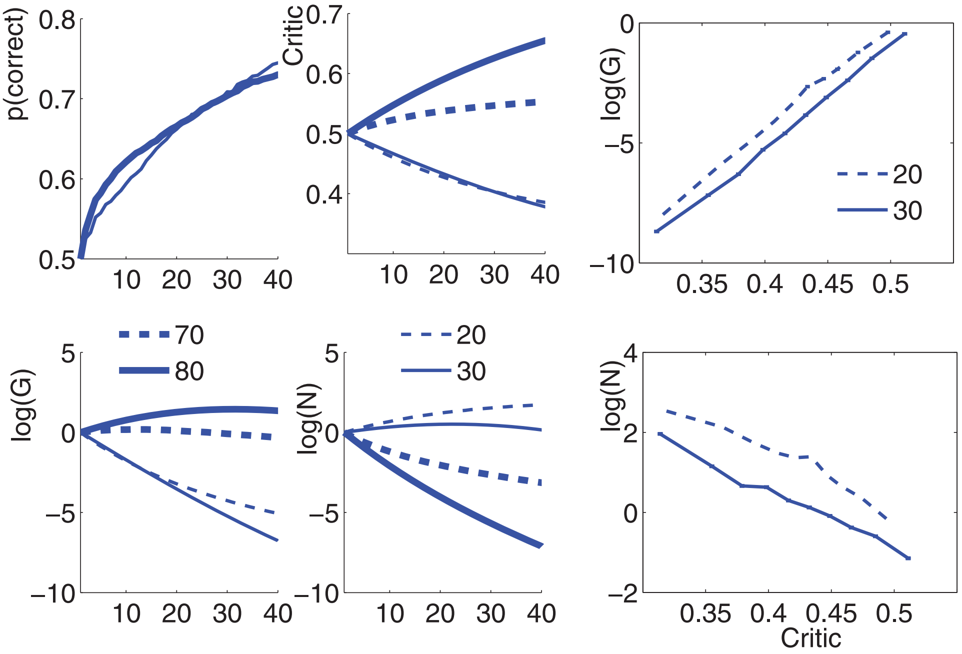

Figure A4. Discrimination learning. Left and middle columns represent opponent actor learning (OpAL) model values across time (top left: probability of choosing the most rewarding option; top middle: critic values; bottom left: logarithm of *G* weight; and bottom middle: logarithm of *N* weight). It is noticeable that critic values for optimal and suboptimal options in the 20–30 case overlap, indicating that a reinforcement learning (RL) model would not know how to separate them. However, *N* weights correctly identify the 20 option as more to be avoided, allowing a better than chance performance. Right columns are diagrams of actor values averaged over decile trials on critic values. For equivalent critic values, *N* weights are higher in average in the 20 case than 30, allowing correct action selection. *G* are also (incorrectly) higher for the 20 case, allowing exploration early on when *G* and *N* weights are comparable and thus contribute equally to the actor but do not hinder choice later due to very low values. See the online article for the color version of this figure.

OpAL. Here we detail the OpAL discrimination learning simulation to explore why OpAL mechanism allows better learning in the 20–30 case (see Figure 11). Investigation of model values shows that, counterintuitively, the average critic value for 20 is very similar to the average critic value for 30 (Figure A4 middle top), reflecting the fact that 20 is sampled less and thus further away from converging to its true value. In an RL model, this would translate into chance performance, since the value estimates drive actor weights in the softmax policy. Here however, actor *N* weights carry more information than critic does: They clearly separate the 20 option from the 30 option (Figure A4 middle bottom), because they tend to exaggerate value representation of worse than expected stimuli via accumulation of multiple negative prediction errors (see main text).

This is very clearly visible in the *phase diagram* (Figure A4 bottom right). Here we separated all trials across time and simulations into 10 quantiles for critic value (separately for each of the four options), and plotted the average *N* value across those trials. We see that for similar critic value, *N* weights are stronger for 20 than for 30. This allows the model to perform above chance even when critic values for the two options are identical. Note that without the Hebbian term in OpAL, the two lines would overlap. Thus, more exploration would be needed to ensure a lower average critic estimate of 20, and better than chance performance. This shows that this result is critically linked to the nonlinear double update rule of OpAL.

Received August 9, 2013
Revision received February 7, 2014
Accepted February 10, 2014 ■