Computational advantages of dopaminergic states for decision making

Alana Jaskir and Michael Frank
Department of Cognitive, Linguistic, and Psychological Sciences
Carney Institute for Brain Science
Brown University
Providence, RI 02912

Abstract

Dopamine's (DA) role in the striatal direct (D1) and indirect (D2) pathways suggests a more complex system than that captured by standard reinforcement learning (RL) models. The Opponent Actor Learning (OpAL) model (Collins & Frank, 2014) presented a more biologically plausible and interactive account, incorporating interactive incentive motivation and learning effects of dopamine in one dual-actor framework. In OpAL, DA modulates not only learning but the influence of each actor on decision making, where the two actors specialize in encoding the benefits and costs of actions (D1 and D2 pathways, respectively). While OpAL accounts for a wide range of DA effects on learning and choice, formal analysis of the normative advantage of allowing the motivational state (the level of dopamine at choice) to be optimized is still needed. We present simulations which suggest a computational benefit to high motivational states in ”rich” environments where all action options have high probability of reward; conversely, lower motivational states have computational benefit in ”lean” environments. We show how online modulation of motivational states according to the environment value or the inference about the appropriate latent state of the environment confers a benefit beyond that afforded by classic RL in learning and risk paradigms. These simulations offer a clue as to the normative function of the biology of RL that differs from the standard model-free RL algorithms in computer science.

Keywords: reinforcement learning; striatal dopamine; cost-benefit tradeoff; risk; motivation

Introduction

Everyday choices involve integrating the known benefits and costs of potential actions. How, then, is it determined when costs matter more than gains? Dopamine (DA) has been shown to play a computational role in both reinforcement learning (RL) and decision making. In RL, phasic firing of midbrain dopamine is widely thought to communicate reward prediction errors (PEs) (Schultz, Dayan, & Montague, 1997). In decision making, dopamine has a performance effect, often characterized in the domain of vigor (how fast to act) (Niv, 2009; Hamid et al., 2015). However, aside from speed of responding, DA also influences the differential weighting of costs and benefits of alternative actions (Collins & Frank, 2014) as observed by DA and striatal D1/D2 modulation of risky choice across species (Zalocusky et al., 2016; Rutledge et al., 2015). Specifically, dopamine inversely modulates two different cell populations of the striatum, which in turn project to different basal ganglia pathways. The medium spiny neurons (MSNs) of the direct pathway express D1 receptors and are strengthened in the presence of DA. Meanwhile, the MSNs of the indirect pathway express D2 receptors and are strengthened in the absence of DA. These mechanisms facilitate opponent reinforcement learning processes in the two populations, and allow the benefits and costs of actions to be differentially weighted during choice (Collins & Frank, 2014).

The Opponent Actor Learning (OpAL) model (Collins & Frank, 2014) accounts for a variety of the D1 and D2 striatal manipulations across species not explained by standard RL models. Based on the biology of the D1 and D2 pathways, OpAL captures interactive learning and choice incentive (performance) effects of dopamine. The model utilizes two actors, one which tracks the benefits of an action and the other the costs of the same action. The *motivational state*, modeled as the amount of dopamine at choice, arbitrates the contribution of each actor on choice. When in a high DA motivational state, D1 striatal activity dominates and choices are made primarily based on the expected benefits of one action over the others and ignoring their relative costs. When in a low DA motivational state, D2 activity dominates and choices are made mostly based on avoiding actions with highest cost. Notably, even with symmetrical learning and balanced motivational states, performance in this two-actor model is more robust to exploration/exploitation tradeoffs across rich and lean environments than that of a standard RL, offering a normative explanation for a dual representation mechanism (Collins & Frank, 2014). However, previous work did not explore the potential added utility of optimizing the level of DA at choice, which could provide a normative explanation for the observed impacts of DA manipulations on risky decision-making.

Here, we have extended OpAL to account for risky decision making by dynamically changing dopamine levels at choice (i.e. motivational state) proportional to the value of the current state. This accounted for findings of increase attractiveness of high-value risky options with the administration of L-DOPA (Figures 1) (Rutledge, Skandali, Dayan, & Dolan, 2015). The model also accounted for individual differences of risk due to effective L-DOPA dosages (data not shown).

Can online modulation of motivational states normatively affect decision making? Is there a computational benefit to one motivational state over another that depends on environmental contingencies? We explored manual manipulations of dopamine at choice in simulated ”rich” and ”lean” environ-

This work is licensed under the Creative Commons Attribution 3.0 Unported License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0

ments. In rich environments, available actions have high probability of reward. In lean environments, available actions have low probability of reward, but optimal performance would still be to choose the action with the highest probability (and not to choose more randomly, as would be predicted from a reduced inverse temperature, or to avoid selecting actions altogether, as might be predicted from classical interpretation of dopamine depletion). We found that a high DA motivational state in rich environments and a low DA motivational state in lean environments improved model performance, in both cases by optimizing choice of the most rewarding action. We next showed that the motivational state could be adapted online as a function of learned reinforcement statistics of the environment, leading to similarly improved model performance. Finally, we assessed performance of a model which assumed structure in how action probabilities are generated across environments (e.g., that the statistics may be bimodal) to infer the motivational state of novel environments. Applying such latent state inference to modulate motivational state further improved performance. These simulations provide a clue as to the normative function of the biology of RL that differs from the standard model-free RL algorithms.

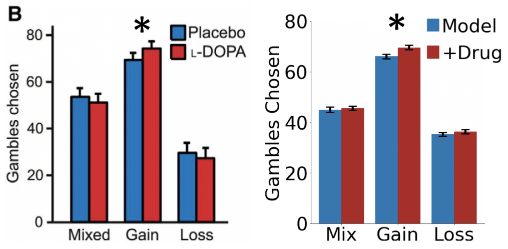


Figure 1: Left: Modified figure from Rutledge et al. (2015). L-DOPA administration selectively increased attractiveness of risky options in gain trials, where choice was between a sure reward and a 50% gamble for a larger reward. Right: Simulation of task using value-modulated OpAL model and selective amplification of high DA at choice, which captured drug effect.

Opponent Actor Learning (OpAL) Overview

OpAL is an actor-critic model. A critic tracks how well the agent does than expected; meanwhile, the actor tracks the value of each action. Rather than computing a single action value as in standard frameworks, OpAL calculates two: one value (G) learns the benefits of an action and the other (N) learns the costs of the same action. G represents the D1 direct (also known as ”Go”) pathway. N represents the D2 indirect (also known as ”NoGo”) pathway. Actors are updated using a three-factor Hebbian rule; this results in anticorrelation but, critically (for the sake of modulation), not symmetry between G and N (Figure 2a).

Critic:

$$V(t+1) = V(t) + \alpha_c \delta(t)$$
$$\delta(t) = r(t) - V(t)$$

Actors:

$$G_a(t+1) = G_a(t) + [\alpha_G G_a(t)] \times \delta(t)$$
$$N_a(t+1) = N_a(t) + [\alpha_N N_a(t)] \times -\delta(t)$$

Policy:

$$Act_a(t) = \beta_G G_a(t) + \beta_N N_a(t)$$
$$\beta_G = \beta(1+\rho) \qquad p(a) = \frac{e^{Act_a(t)}}{\sum_i Act_i(t)}$$
$$\beta_N = \beta(1-\rho)$$

Throughout the duration of this paper, we set $\alpha_G = \alpha_N$ for simplicity. $V(t)$ can represent either a state-action pair or state value.

Dopamine is represented in this model in two ways. **The *motivational state*, or the level of dopamine at choice, is represented by ρ and affects the expression of previously learned values for decision making**. As shown in the Policy Equations, ρ arbitrates the weighting of G (benefits) and N (costs) values. With higher DA, i.e. positive ρ, choices are made predominately on benefits. With lower DA, i.e. negative ρ, choices are made predominately on costs. Secondly, we have the standard prediction error, $\delta(t)$, which is widely believed to be encoded in phasic DA. Importantly, positive and negative prediction errors contribute to learning for *both* actors, but with oppositive effects; it is the nonlinearity in the update (three-factor Hebbian rule) which gives rise to the benefit/cost specialization of the actors.

Manual ρ modulation

Is there a computational benefit to different motivational states? We tested this by comparing a model optimized in rich and lean environments with $\rho = 0$ and the same model with modulated ρ (all other parameters held constant).

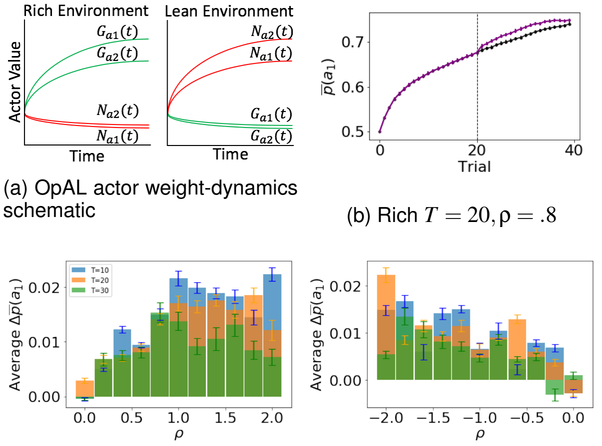
In rich environments, models learned to discriminate between two actions with reward probabilities: $p_{a1}(r) = .8$ and $p_{a2}(r) = .7$. In lean environments, models learned to discriminate between actions: $p_{a1}(r) = .3$ and $p_{a2}(r) = .2$.

Both optimized models and modulated models learned over 40 trials. Until a threshold of trials learned, T, the modulated models and optimized models had equivalent parameters ($\rho = 0$). Each modulated model was paired with an optimized model and was forced to select the same action and received the same feedback as the paired optimized model until time T. After a threshold of trials learned T, ρ was set to some fixed value, $\rho \in [-2, 2]$, for the remainder of learning for the modulated models. Parameter values as optimized in Collins and Frank (2014): $\alpha_c = .035, \alpha_G = \alpha_N = .98, \beta = 1.5, V_a(0) = 0.5, G_a(0) = N_a(0) = 1$.

Results

We found that increasing ρ (i.e. dopamine at choice) in rich environments improved performance (Figure 2c) while decreasing ρ impaired performance (not shown). Conversely, decreasing ρ in lean environments improved performance (Figure 2d) while increasing ρ impaired performance (not shown).

For the rich environment, the average softmax probability of selecting the most rewarding action after trial $T = 20$ increased when modulated models are in a positive ($\rho = .8$) DA



(a) OpAL actor weight-dynamics schematic

(b) Rich $T = 20, \rho = .8$

(c) Rich $\rho \geq 0$

(d) Lean $\rho \leq 0$

Figure 2: (a) In rich environments, G values are more differentiated. In lean environments, N values are more differentiated. (b) Rich environment example performance curves for optimized model (black) and modulated model (purple) with $\rho = .8$ after $T = 20$ trials. Performance was measured as the probability of selecting the most rewarding action (a_1) according to the softmax policy. Average performance was calculated over 10,000 modulated and unmodulated models each. (c) Average difference between performance curves after trial T for different T and $\rho \geq 0$ for rich environment. (d) Average difference between performance curves across time points for different T and $\rho \leq 0$ for lean environment. All error bars represent SEM.

motivational state (Figure 2b). Generally, we found that the advantage of the modulated model in these performance curves after a given time T increased with larger positive values of ρ and with earlier times T of modulation (Figure 2c). However, performance of the modulated model progressively decreased with increasingly negative values of ρ (not shown). In the lean environment, we conversely found that larger negative values of ρ increased performance of the modulated model (Figure 2d) while increasingly positive values of ρ impaired average performance (not shown).

Discussion

These simulations suggest it is computationally beneficial to be in a higher motivational state in environments where all action options have high probability of reward; alternatively, it is beneficial to be in a lower motivational state when action options yield low probability of reward. These results can be interpreted intuitively in terms of achieving the best discrimination among options (Figure 2a). In rich environments, more differentiation exists in the G weights of the actions. It then follows that it would improve performance to take G values more into consideration by increasing ρ. On the other hand, in lean environments, more differentiation exists in the N weights of the actions. It then follows that it would improve performance to take N values more into consideration by decreasing ρ.

Online ρ modulation by state value

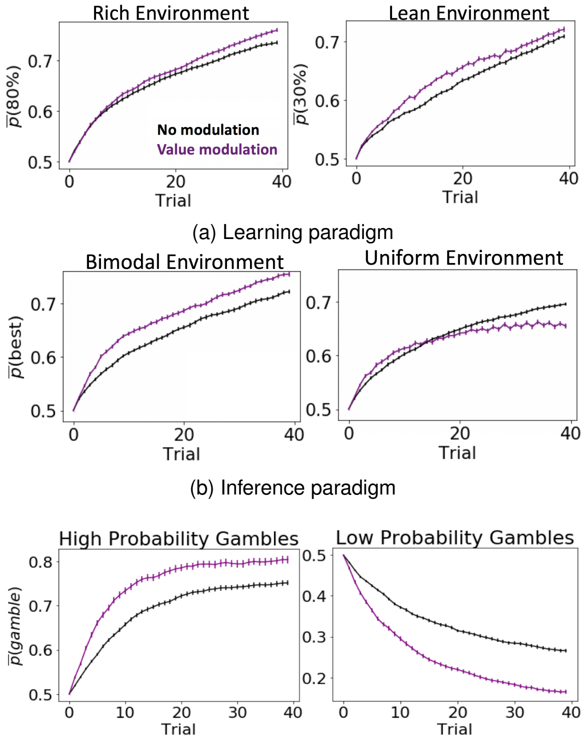
Is there an online way to determine ρ and leverage this computational advantage of motivational states? We investigated whether state value could be used to dynamically modify the dopamine levels at choice (ρ) in order to improve performance compared to a model without such modulation ($\rho = 0$). We found that that modulating ρ online by learned or inferred state-value could improve performance in separate learning, inference, and risk paradigms.

Learning: We trained models in reward rich (80% vs. 70%) and lean (30% vs. 20%) states. In addition to tracking actor and critic values for the rich and lean environment, the value of an state $V_s(t)$ was calculated by averaging over the two critic action values for that state. ρ was set to the trial-by-trial $V_s(t)$ scaled by some constant. ***Inference:*** If an agent assumed or has learned some statistical structure about its world (e.g. there exist rich states and lean states), would inferring the latent statistical structure to determine ρ be beneficial to performance in novel contexts? We simulated 10,000 different states from two different types of generative distributions. Each state had two actions. In the Uniform environment, average $p(r)$ of the two actions was drawn uniformly from $p(r) \sim U(.2,.8)$, where $p_{a1}(r) = p(r) + .05, p_{a2} = p(r) - .05$. In the Bidmodal environment, the state was either high (average $p(r) = .8 + \sim U(.05,-.05)$) or low (average $p(r) = .2 + \sim U(.05,-.05)$) and $p_{a1}(r) = p(r) + .05, p_{a2} = p(r) - .05$. Using Bayes rule, the modulated model inferred the state (rich or lean) given reward history. Motivational state ρ was determined trial-by-trial by averaging the explicitly assumed expected value of rich or lean states weighted by the calculated probability that a state was rich or lean. ***Risk:*** As previously mentioned, value modulation of motivational states allowed OpAL to account for risky decision making in Rutledge et al. (2015) (Figure 1). Could online modulation be helpful for inferring when it is advantageous to select a risky option in the long run? Models selected between a sure reward and a gamble of twice the value with unknown stable probability; gamble reward was encoded relative to the sure thing. In high probability gamble states, the probability of reward was drawn uniformly above 50%; in low probability states, probability of reward for the gamble was drawn uniformly below 50%. Models were presented with the same gamble for 40 trials. The critic (tracking the value of selecting the gamble) modulates ρ.

Absence of reward for all paradigms were encoded as a cost, $r = -1$. Parameters: $\alpha_c = .05, \alpha_G = \alpha_N = .3$ or .1 (risk paradigm),$\beta = 1.5$

Results

We found that state value could be used to modulate motivational state ρ online, outperforming a model with matched parameters but without ρ modulation. In the learning paradigm (Figure 3a) like in the manual modulation, higher state value corresponded to higher dopamine levels at choice in rich states, which therefore improved decision making. Alternatively, lower state value, corresponding to lower dopamine lev-



(a) Learning paradigm

(b) Inference paradigm

(c) Risk paradigm

Figure 3: Model performance measured as the probability of selecting the most rewarding action according to the softmax policy. Average performance was calculated over 10,000 value modulated and unmodulated ($\rho = 0$) models. Error bars represent SEM.

els, in the lean states improved performance.

We found that in the Bimodal case, a model which inferred the latent motivational state of the environment, outperformed the no modulation model. As a contrast, the value model had more variable performance in the Uniform case, suggesting false structural assumptions impaired performance (Figure 3b). While this must be investigated further, likely environments approaching $p(r) = .5$ had a higher probability of inferring a ρ in the opposite sign than what would be beneficial. Further work will compare these models with optimized parameters for the task.

Finally, using the value learned from selecting a gamble to modulate ρ online also improved performance in comparison to a non-modulated model (Figure 3c). In states with high probability ($> 50\%$), value modulation helped the model infer that the gamble was advantageous. In low probability gambles, value modulation aided in avoiding the gamble.

Discussion

These results suggest that environment value could be used to modulate motivational states. This work intersects with studies which show that dopamine during anticipation of choice reflects a discounted value function (Hamid et al., 2015). Furthermore, these results offer a normative explanation for findings that pharmalogical manipulations, dopamine depletion, and optogenetic stimulations of D1 and D2 pathways affect cost/benefit choice and effort. In this perspective, the brain treats increases or decreases in dopamine as signaling presence in a richer or leaner state. Hence, a dopamine depleted animal would focus on costs of actions and dopamine increases would increase attractiveness of risky actions (Rutledge et al., 2015).

General Discussion

These findings suggest that optimizing the motivational state, or the level of dopamine at choice, provides computational benefits for performance over and above the previously reported finding that OpAL is more robust to contingencies than standard RL even without modulation of motivational state. Specifically, we found a benefit of higher motivational states when in rich environments where all action options have high probability of reward; alternatively, lower motivational states aid performance in lean environments. The appropriate level of dopamine can be modulated online according to the value of the current environment or through inferring which environment type the agent is in given a known structure. These offer normative explanations for dopamine's affect on behavioral cost/benefit tradeoff unaccounted for by standard-RL, as well as presenting a functional role for dynamically dialing the relative weighting of benefits and costs to best capitalize on the representations of action values in the D1/D2 pathways to improve decision making. Follow-up studies will continue to explore the relationship between latent structure of environments and inference of motivational states.

Acknowledgments

Support for this work comes from NIMH R01 MH080066-01 and NSF grant 1460604

References

Collins, A. G. E., & Frank, M. J. (2014). Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychological Review*. doi: 10.1037/a0037015

Hamid, A. A., Pettibone, J. R., Mabrouk, O. S., Hetrick, V. L., Schmidt, R., Vander Weele, C. M., ... Berke, J. D. (2015). Mesolimbic dopamine signals the value of work. *Nature Neuroscience*. doi: 10.1038/nn.4173

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*. doi: 10.1016/j.jmp.2008.12.005

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2015). Dopaminergic Modulation of Decision Making and Subjective Well-Being. *Journal of Neuroscience*. doi: 10.1523/jneurosci.0702-15.2015

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*. doi: 10.1126/science.275.5306.1593