Cognitive Science Honors Thesis

A Computational Account of Sensory Prediction Error Gating in Reinforcement Learning Models

Matthew Boggess

December, 2015

Cognition and Action Lab, UC Berkeley

Advisors:

Matthew Crossley, Richard Ivry

Faculty Reader:

Ming Hsu

Abstract

A successful return in tennis requires a tennis player first to determine where best to place her return and then to correctly execute her swing. If she makes an errant return, she now faces a credit assignment problem: Should this negative outcome be attributed to poor shot selection or to an error in motor execution? McDougle et al. propose a solution to this problem when the source of the error is the motor system. They posit that motor errors are communicated to the decision-making system whereby they gate learning in order to prevent the undesired negative reinforcement of the chosen action. This gating hypothesis was motivated by recent anatomical evidence that the cerebellum — a crucial node in a network widely thought to process motor execution errors — sends direct subcortical projections to the basal ganglia — a crucial node in a network widely thought to drive reinforcement learning and decision-making. In McDougle et al.'s gating model, motor execution errors scale learning rates in a temporal difference (TD) reinforcement learning model of decision-making. However, the most prominent attempts to link the basal ganglia to reinforcement learning models have instead suggested that actor-critic (AC) models may be more appropriate models of basal ganglia anatomy. In the present study, we investigate the gating hypothesis from the perspective of AC models. We find that AC models can account for McDougle et al.'s behavioral results, but that gating is necessary for them to do so. Additionally, we find that simultaneous gating of the actor and critic is the only AC gating model that can account for subject data, but that AC models do not account for subject data better than TD models. We conclude with a discussion of the biological plausibility of the proposed gating mechanism from the perspective of the AC gating model.

Introduction

When returning an opponent's shot in tennis, a tennis player must first determine where best to place her return and then correctly execute her swing to produce a successful return. Imagine she decides to return the shot down the line, but then hits the ball into the net because of an error in her swing. Clearly this is not the desired outcome, but now she faces a credit assignment problem: Should this negative outcome be attributed to a poor shot selection or to an error in motor execution? In order for the brain to resolve this credit assignment problem, some process must allow the decision-making and motor control systems to communicate. Despite this problem, decision-making and motor control have traditionally been studied and modeled independently.

Two neural structures thought to be important for motor control and decision-making respectively are the cerebellum and the basal ganglia. The cerebellum has widely been theorized as a supervised learning system (Doya, 1999; Wolpert et al., 1998). The system learns to predict the sensory consequences of efferent motor commands, which are used to correct ongoing actions. Sensory prediction errors (SPE) generated from the difference between these predictions and the actual observed sensory consequences serve as the teaching signal. In contrast, the basal ganglia, a neural structure important for action selection, has been widely theorized as a reinforcement learning system (Doya, 1999). The system learns to repeat actions that lead to positive outcomes and avoid actions that lead to negative outcomes. Reward prediction errors (RPE's), which are the difference between the predicted reward and the actual reward obtained from performing an action, serve as the reinforcement signal.

Acting in isolation, neither of these two systems can resolve the credit assignment problem. In the case of our tennis player, both an SPE and a negative RPE are generated. The SPE will be used to appropriately correct the motor plan to avoid future mistakes in the swing, but without any knowledge of the SPE, the RPE will negatively reinforce the chosen shot. Ideally, these two systems would instead directly communicate to prevent this negative reinforcement. Past theories have considered interactions between the cerebellum and basal ganglia. Most prominently, Houk and colleagues developed distributed processing module theory, which treats the cerebral cortex, cerebellum, and basal ganglia as distinct, yet connected modules, and have shown how this framework accounts for a variety of behavioral and physiological findings (Houk, 2005; Houk & Wise, 1995). However, no prior effort has considered the credit assignment problem addressed in this study.

McDougle et al. hypothesize that the credit assignment problem can be solved if errors in motor execution gate learning in the decision-making system (Fig. 1A). This gating prevents negative reinforcement of the chosen action when the negative reward is attributed to be a consequence of a motor error. They cite anatomical projections from the deep cerebellar nuclei of the cerebellum to the striatum of the basal ganglia as providing the neural substrate for this gating mechanism (Hoshi et al., 2005). This pathway could allow SPE’s generated in the cerebellum to be communicated to the basal ganglia and used to prevent negative reinforcement of the chosen action.

To test this hypothesis, McDougle et al. used a classic decision-making behavioral design known as the “n-armed bandit problem”. In this task, subjects repeatedly selected between two slot machines (n = 2) with different reward schedules unknown to the subjects. The subjects' goal was to maximize the total number of points received. Consistent with previous findings, subjects were not only able to reliably track the value of the slot machines, but exhibited a risk averse bias, selecting the slot machine that paid out points more often despite the options having equal expected value (Daw et al., 2006; Kahneman & Tversky, 1979). To bring the motor system into this decision-making process, McDougle et al. created a novel variant in which participants selected slot machines by making simple reaching movements to the desired slot machine instead of pressing keyboard keys. Though the schedule of reward was kept identical across conditions, subjects now believed the lack of payout from the selected slot machine was a consequence of an error in the reach, as opposed to being a property of the slot machine itself. In this new condition, subjects completely reduced their risk averse bias, instead displaying a risk seeking bias. McDougle et al. also showed that this effect was independent of subjects' sense of control over the action, suggesting an implicit mechanism. Furthermore, patients with cerebellar ataxia did not reduce their risk averse bias, suggesting a critical role for the cerebellum.

McDougle et al. used a reinforcement learning model known as the temporal difference (TD) model to argue that the gating hypothesis explains the observed shift in risk bias (Sutton & Barto, 1998). The TD model has become widely utilized in neuroscience due to the similarity between midbrain dopaminergic activity (DA) and the TD error signal (Schultz et al., 1997). TD models maintain a value function that maps actions to value estimates (in this design, one value estimate for each slot machine). The value function for an action is updated proportional to the resulting RPE after performing that action. McDougle et al. showed that a variant of this classic TD model with separate learning rates for trials with execution errors (miss trials) and without execution errors (hit trials) fit subject data significantly better than the classic TD algorithm. The dual learning rates allow for the realization of gating by having a smaller learning rate for miss trials compared to hit trials. Consistent with the gating hypothesis, model fits to subject data in McDougle et al. revealed that the salient difference between the keyboard and reaching conditions was a decrease in the miss trial learning rate for the reaching condition (Fig. 1B).

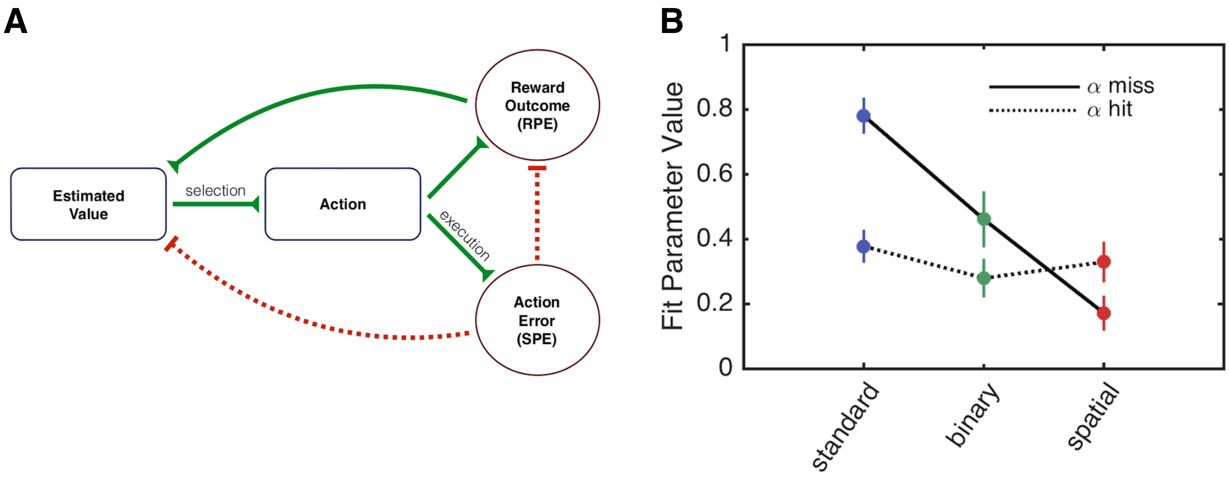


Figure 1: *Gating model from McDougle et al., in review*. (**A**) Schematic of gating hypothesis from McDougle et al., in review. Gating could occur either by abolishing the RPE or preventing the value estimate update. Gating in the TD model occurs by reducing value estimate updates on miss trials (SPE's present). (**B**) Best fit parameters for McDougle et al.'s gating model across conditions. The salient difference between the standard and spatial conditions is the reduction of α_{miss} in the spatial condition. This is consistent with reduced value estimate updates in the presence of execution errors in line with the gating hypothesis.

While the classic TD model is the most commonly used reinforcement learning model for modeling behavior, the actor-critic (AC) model is the prominent reinforcement learning model attempting to account for basal ganglia anatomy (Sutton & Barto, 1998). Unlike the TD models, AC models separate the value function and decision policy into the critic and actor respectively. The critic is responsible for generating an error signal that is used to both update its own value estimates and to update the decision policy of the actor (Barto, 1995). In the models presented here, the critic is identical to the classic TD model. However, unlike the classic TD model, RPE's generated by the critic are also used to update the actor's decision policy, and the response made by the model is determined by the actor, not directly by the value estimates.

Though the AC model does exhibit some theoretical advantages over the classic TD model (Sutton & Barto, 1998), its primary appeal is its alignment with basal ganglia anatomy and physiology. As discussed previously with the classic TD model, the prediction error signal generated by the critic is strikingly similar to midbrain dopaminergic activity (Schultz et al., 1997). However, unlike the classic TD model, the AC model is also motivated by the similarity between DA-dependent plasticity in the striatum and learning guided by a prediction error in the actor (Calabresi et al., 2000; Wickens et al., 1996). Consequently, multiple AC models of the basal ganglia have been developed with the striatum playing a key role (Brown et al., 2000; Collins & Frank, 2014; Houk et al., 1995; Suri & Schultz, 1998; Suri et al., 2001).

The central role of the striatum in AC models of the basal ganglia is consistent with the anatomical motivation for the gating hypothesis. Since the striatum is both a terminal of the cerebellar-basal ganglia projections as well as the central site for learning in the AC model of the basal ganglia, SPE's along this pathway appear appropriately located to have a significant effect on learning. AC models also raise an important question regarding how the gating hypothesis fits into a biologically plausible framework: Do execution errors gate the actor (response policy updates), the critic (value estimate updates), or both? Finally, an alternative possibility is that the AC architecture completely negates the need for gating.

In order to address these questions, it is necessary to characterize the full range of risk biases the models are capable of producing. Niv et al. showed that risk aversion emerges naturally as a consequence of optimal reinforcement learning and does not require additional assumptions such as a nonlinear subjective utility curve (Niv et al., 2002). Past work has also shown that risk seeking biases are achievable when choosing between a certain and uncertain option of equal expected value (Denrell, 2007; March, 1996). However, these analyses do not extend to the case of choosing between two uncertain options. Thus, it remains unclear what biases the considered models are capable of producing and whether gating is necessary to produce the behavior observed in subjects. To address these unknowns, we use a technique known as parameter space partitioning (PSP), which allows for the characterization of the full range of behaviors the models are capable of producing (Pitt et al., 2006; Pitt et al., 2008).

Previewing our results, we find that the AC architecture can account for McDougle et al.'s behavioral results but that gating is necessary because the classic AC model cannot produce a risk seeking bias. Additionally, we find that simultaneous gating of the actor and critic is the only gating model that produces an improvement in fit compared to the classic model, but AC models do not provide superior fits to TD models. Finally, we find that other parameters besides gating can shift risk biases that themselves have behavioral implications.

Methods

Behavioral Design (McDougle et al., in review)

All models were simulated on data obtained from the behavioral tasks of McDougle et al., in review. This task consists of selecting between two targets that either gave points between 1 and 100 (hit trial) or gave 0 points (miss trial), according to predetermined payoff and hit probability curves on each trial (Fig. 2A & B). Subjects complete a total of 600 trials. In the standard condition, subjects select between the two targets by pressing either the left or right arrow keys (Fig. 2C). In the spatial and binary conditions, subjects select between the two targets by making center out reaches to the left or right (Fig. 2D). In both the spatial and binary conditions, subjects are instructed that miss trials are a consequence of errors in their reach. In the standard condition, subjects are instructed that miss trials are a property of the target. Spatial and binary conditions differ in that subjects receive feedback correlated with their reach in the spatial condition, but they receive no feedback as to the accuracy of their reach in the binary condition. Unbeknownst to subjects, feedback is manipulated in the binary and spatial conditions such that the underlying schedule of rewards is identical across all conditions.

Hit and payoff functions were constructed such that the expected value of the two targets is equal. A “safe” target delivers more hits at lower payoffs, and a “risky” target delivers fewer hits at higher payoffs. Each function is the superposition of low-frequency pseudo-sinusoids and high-frequency Gaussian noise. Hit and payoff functions are perfectly inversely correlated within targets and are perfectly correlated between targets. A safe choice is made whenever the selected target has a hit probability greater than one half for that trial and vice versa for a risky choice.

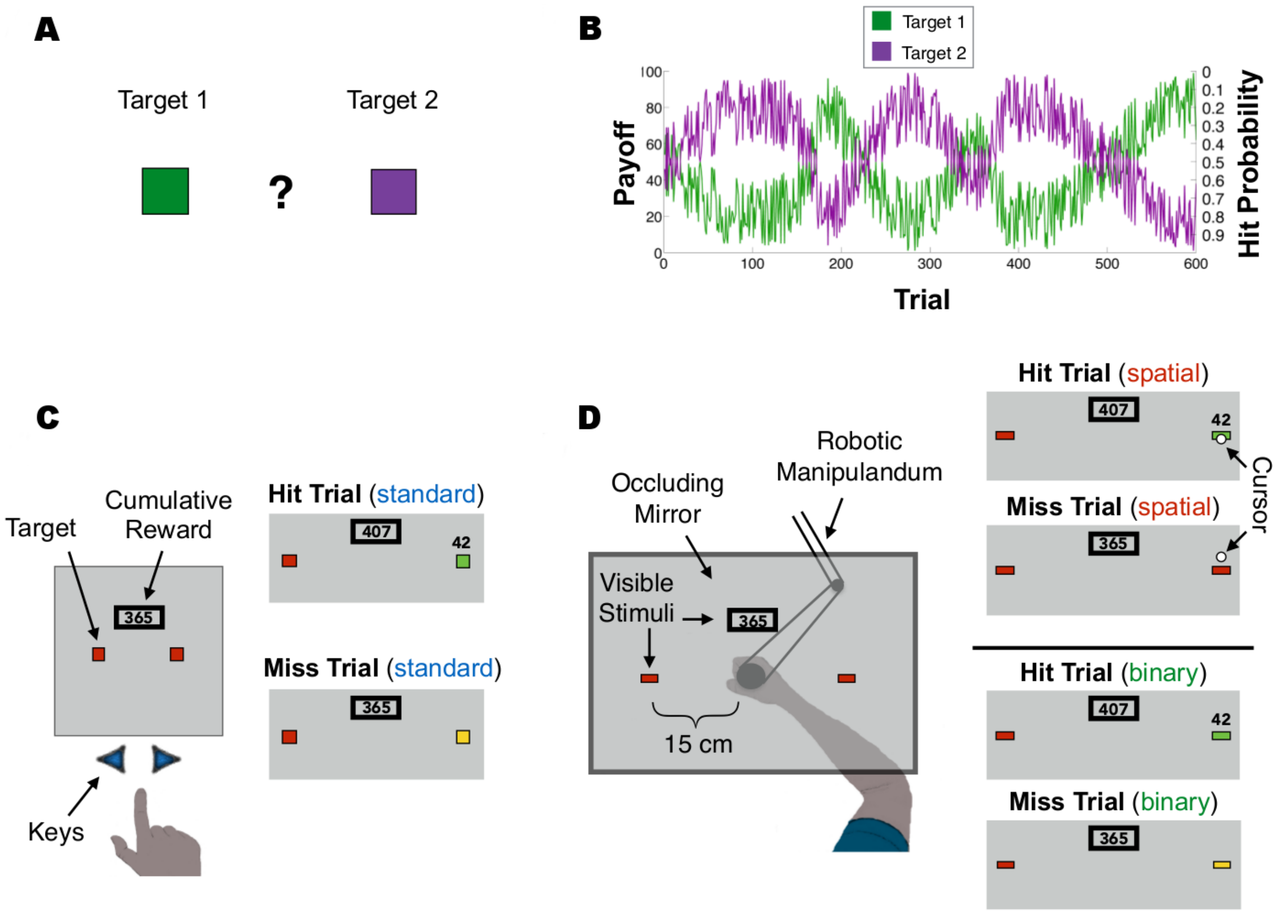


Figure 2: *Design from McDougle et al., in review*. (**A**) Participants performed a 2-armed "bandit task", choosing between two targets to maximize monetary payoff. (**B**) Two reflected, noisy sinusoids defined the payoff value (left axis) and probability of reward ("hit", right axis, inverted) for the targets. (**C**) In the Standard condition participants selected targets by pressing the left or right arrow keys on a keyboard. Example "hit" and "miss" trials are shown on the right. (**D**) In the Spatial and Binary conditions, participants reached to the selected target using a robotic manipulandum. Vision of the hand was occluded. In the Spatial condition, a small cursor appeared after the hand passed the target On "hit" trials, the cursor overlapped with the target; on "miss" trials, the cursor appeared outside the target Feedback in the Binary condition matched the Standard condition.

Temporal Difference Model Architectures

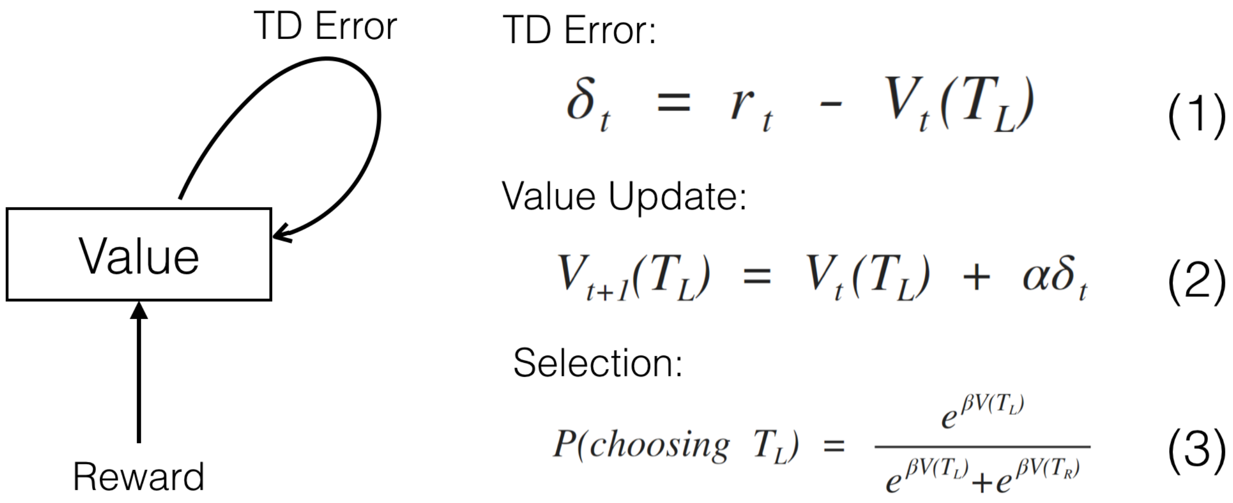
In addition to the AC models described in the next section, we also tested two TD models. The first TD algorithm was the classic temporal difference (TD) algorithm (Sutton & Barto, 1998). This algorithm maintains separate value estimates for each target, which are updated after each choice by a prediction error. The prediction error is simply the difference between the observed reward and the predicted reward of the chosen target on that trial (Fig. 3, Eq. 1). The observed reward is taken from the schedule of rewards shown in Fig. 2B. The predicted reward is simply the value estimate for the chosen target on that trial. Prediction errors update value estimates according to Fig. 3, Eq. 2 and only the value estimate for the target that was chosen is updated. The softmax function is used to probabilistically select responses on each trial based on the current value estimates of the targets (Fig. 3, Eq. 3).

The two parameters of interest in this model are α and β. α controls the degree to which the model incorporates its most recent observation into its future predictions. β (also known as the inverse temperature) controls the model's tradeoff between exploration and exploitation. Higher values of β make the target with the higher value much more likely to be chosen (exploitation) while lower values make choosing the lower valued target more probable (exploration).

We also examined a second TD model designed to capture the idea that errors in motor execution gate reinforcement learning (the gating hypothesis from the introduction). It is identical to the classic TD model except that miss trials follow a different value estimate update:

$$\text{TD Gating:} \quad V_{t+1}(T_L) = V_t(T_L) + \alpha_{gate} * \alpha \delta_t \qquad (4)$$

The gating rate (α_{gate} in Eq. 4) determines how much less the value will be updated on a miss trial compared to a hit trial (the value update for hit trials is the same as in Eq. 5). For $\alpha_{gate} = 1$, the TD gating algorithm reduces to the classic TD algorithm. As α_{gate} approaches zero, the feedback from miss trials is discounted more and more. Note that gating in these models is not sensitive to the size of the motor error, but instead is present in equal magnitude on all miss trials.



TD Error:

$$\delta_t = r_t - V_t(T_L) \tag{1}$$

Value Update:

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

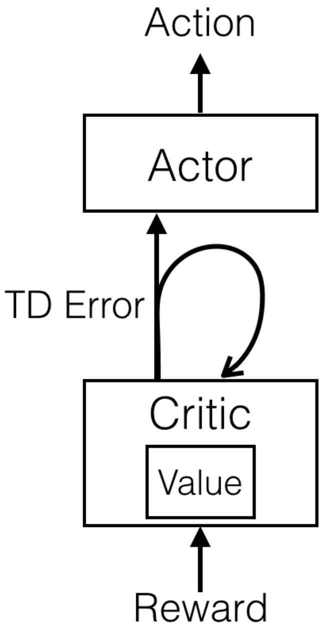
Selection:

$$P(choosing\ T_L) = \frac{e^{\beta V(T_L)}}{e^{\beta V(T_L)} + e^{\beta V(T_R)}} \tag{3}$$

Figure 3: *TD Model*. (**Left**) Schematic of TD model. (**Right**) TD model Equations: (**1**) Prediction error (δ_t) computation for a trial where the left target was chosen. The prediction error is the difference between the received reward after choosing the left target (r_t) and the current value estimate of the left target (V_t). (**2)** Value function update after choosing the left target. The value estimate for the next trial is updated proportional to the prediction error according to the learning rate (α). Both value estimates are initialized to 0. (**3**) The model selects a target probabilistically according to the softmax function over the target's value estimates.

Actor-Critic Model Architectures

AC models split the selection policy and the value estimation into separate modules called the actor and the critic respectively (Sutton & Barto, 1998). The critic maintains value estimates for the two targets and learns via the same prediction error mechanisms described for the classic TD model (Fig. 4, Eq. 6). The actor maintains the response policy, which operates on different estimates than that represented by the critic. These are updated after each choice via the same prediction error generated and used by the critic (Fig. 4, Eqs. 5 & 7). This highlights a major difference between the classic TD and AC models, since the response policy in the AC models is no longer a direct function of the value estimates.



Actor Update:

$$W_{t+1}(T_L) = W_t(T_L) + \alpha_{actor}\delta_t \qquad (5)$$

Critic Update:

$$V_{t+1}(T_L) = V_t(T_L) + \alpha_{critic}\delta_t \qquad (6)$$

Selection:

$$P(choosing\ \ T_L) = \frac{e^{\beta W(T_L)}}{e^{\beta W(T_L)} + e^{\beta W(T_R)}} \qquad (7)$$

Figure 4: *Actor-Critic Model*. (**Left**) Schematic of AC model. (**Right**) AC model Equations: (**1**) Actor update after choosing the left target. The weight for the left target is updated proportional to the prediction error according to the actor learning rate (α_{actor}). Both weights are initialized to 0. (**2)** Critic update after choosing the left target. The value estimate for the left target is updated proportional to the prediction error according to the critic learning rate (α_{critic}). Both value estimates are initialized to 0. (**3**) The model selects a target probabilistically according to the softmax function over the actor target weights.

Due to the separate learning rates for the actor (α_{actor}) and critic (α_{critic}), the separation of the actor and the critic affords the AC model more flexibility. α_{critic} controls the volatility of the prediction error signal. High and low values of α_{critic} lead to sustained prediction errors due to the susceptibility to local noise and inability to track the underlying value respectively. However, unlike the TD models, the chosen action is no longer directly dependent on these value estimates as α_{actor} controls how dynamically the response policy shifts in response to current prediction errors. Note that this AC model reduces to the TD classic model when $\alpha_{actor} = \alpha_{critic}$.

The AC model also produces new forms of gating. Specifically both the actor and the critic can now be subject to gating:

$$\text{Critic Gating: } V_{t+1}(T_L) = V_t(T_L) + \alpha_{gate} * \alpha_{critic} \delta_t \qquad (8)$$

$$\text{Actor Gating: } W_{t+1}(T_L) = W_t(T_L) + \alpha_{gate} * \alpha_{actor} \delta_t \qquad (9)$$

Here we consider three gating variants along with the classic (no gating) AC model: actor gating, critic gating, and dual gating. Critic gating reduces updates of the critic value estimates on miss trials (Eq. 8), actor gating reduces updates of the actor weights on miss trials (Eq. 9), and dual gating reduces updates in both the actor and the critic on miss trials at a shared rate. We chose the same gating rate for both the actor and critic in the dual gating model in order to reduce complexity. Gating in the actor and critic could also conceivably occur at different rates, but we did not investigate such a model here. Note that the dual gating model reduces to the TD gating model when the actor and critic learning rates are equal.

Model Characterization - Parameter Space Partitioning

A primary goal of the analysis in this study is to determine if gating is a plausible mechanism for generating the risk seeking behavior that was observed in the spatial condition. In order to do so, it is necessary to characterize the full range of behaviors these models are capable of producing. Parameter space partitioning (PSP) is used to exhaustively and efficiently search the parameter space of a model in order to map out all of the distinct qualitatively different behaviors it can produce (Pitt et al., 2006; Pitt et al., 2008).

PSP partitions the parameter space into regions that produce qualitatively different data patterns defined by the experimenter. PSP uses a Monte Carlo Markov Chain search algorithm to efficiently search the parameter space of the model. Once the mapping is complete, PSP outputs the estimated volume of each of these regions. Patterns for the models were defined based on the risk bias of the models (see Table 1 for a summary of patterns). The risk bias is the probability that the model would choose the risky target (Eq. 10).

$$P(Risky)=\frac{\#\ \ Risky\ \ Choices}{\#\ \ Trials} \qquad (10)$$

Higher bias values imply more risk seeking behavior and lower values imply more risk averse behavior. Note that risk seeking behavior discussed here means that subjects show a bias towards picking the riskier target. It does not necessarily mean they actually have an explicit preference for riskier options, but instead it arises from some other mechanism such as gating. Two thresholds were chosen in order to initially map each bias into one of three patterns: risk seeking, risk neutral, and risk averse.

Risk biases ≥ 0.6 were classified as risk seeking, risk biases ≤ 0.4 as risk averse, and risk biases between 0.4 and 0.6 as risk neutral. The thresholds 0.4 and 0.6 were chosen for two reasons. First, they are both less biased than the subject averages in McDougle et al. and thus the full range of biases should extend sufficiently beyond these thresholds. Second, parameter settings that produced the smallest biases in each pattern still produced reliable tracking of either the safe or risky target respectively. Two additional patterns were added to account for the fact that perseveration to either of the targets will produce a substantial bias despite this behavior clearly not reflecting genuine risk seeking or aversion. Perseveration was detected whenever the average model choice behavior reached to one target for more than 100 consecutive trials and was split into two patterns depending on which target was selected.

Pattern	Definition
Risky Perseverating	More than 100 consecutive selections of the riskier target.
Risk Seeking	Bias ≥ 0.6
Risk Neutral	0.4 < Bias < 0.6
Risk Averse	Bias ≤ 0.4
Safe Perseverating	More than 100 consecutive selections of the safer target.

Table 1: Summary of the five patterns used for PSP. The bias is calculated as in Eq. 10.

A MATLAB implementation of PSP was obtained from the website of J. I. Ayung (**http://faculty.psy.ohio-state.edu/myung/personal/psp.html**). All parameters were allowed to vary from 0 to 1. Learning rates were capped at one to remain consistent with the analysis in McDougle et al. Gating rates were capped at one to reflect the desired property of gating to solely reduce learning on miss trials. β was capped at one for greater computational tractability because the range of the inputs to softmax were large enough that higher values of β generally produced similar behavior.

PSP makes three key assumptions about the model in order for it to work properly:

1. Data patterns occupy a single continuous region in parameter space.
2. Data pattern regions are contiguous, i.e. there are no undefined regions of parameter space between defined regions.
3. Model output is stationary, i.e. identical parameter settings deterministically produce the same output data pattern.

Assumptions 1 and 2 are difficult to confirm directly. However, one can effectively check for satisfaction of these assumptions by running the PSP algorithm multiple times. If either of these assumptions are violated, there will be a lack of consistency between volume estimates of parameter regions across runs. Consequently, each of the PSP analyses were run 10 times for each model. The volume estimates were highly robust across the different runs as quantified by the standard error of the mean. Assumption 3 is violated at the level of a single simulation since the models are probabilistic. Thus, each model was simulated 100 times and the average was fed into PSP. Variation in the bias after averaging 100 simulations was minimal resulting in consistent data patterns across different simulations.

Model Comparison - Bayesian Information Criterion

PSP is simulation based and does not take subject behavior into account. In order to quantify how well the models fit McDougle et al.'s subject data, we used the Bayesian Information Criterion (BIC, Eq. 11, Schwarz et al., 1978).

$$BIC = -2ln(L) + kln(n) \tag{11}$$

L is the maximum likelihood of the model fit, k is the number of free parameters of the model, and n is the number of data points (600 trials). Model fits were obtained using the MATLAB fmincon function to find parameter values that minimized the negative log likelihood of all subjects' choices simultaneously. All model parameters were optimized individually for each subject and were allowed to vary between 0 and 1 during the optimization process (same range as used in PSP). Lower BIC values correspond to better fits to subject data.

Results

Gating is necessary to produce risk seeking behavior

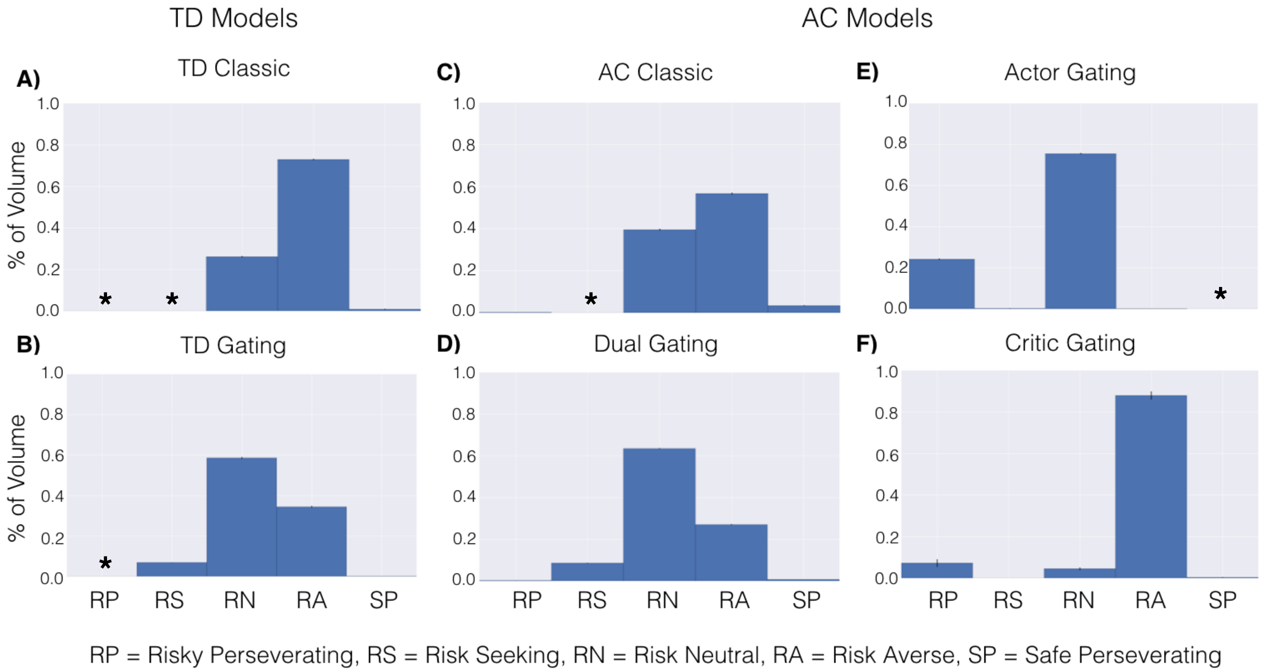


Figure 5: *PSP Pattern Volume Estimates*. Volumes of patterns are averaged across ten separate PSP runs (error bars are s.e.m.) and normalized by the total volume of the model's parameter space. * denotes patterns with zero volume indicating the model is incapable of producing behavior represented by that pattern. Although some volumes without stars may appear to be zero, these patterns simply had tiny portions of parameter space that produced this behavior. (**A**) TD classic model volumes. (**B**) TD gating model volumes. (**C**) AC classic model volumes. (**D**) AC dual gating model volumes. (**E**) AC actor gating model volumes. (**F**) AC critic gating model volumes.

Volume estimates for the different patterns for each model produced by PSP can be found in Fig. 5. To be consistent with McDougle et al.'s behavioral results requires that a given model can produce both risk seeking and risk averse behavior. As can be seen in Fig. 5, all six models are capable of producing risk averse behavior. However, the classic TD and AC models are not capable of producing risk seeking behavior (Fig. 5A & C). Thus, gating is necessary to produce the risk seeking behavior observed in the spatial condition. This is consistent with McDougle et al.'s finding that the TD gating model fit human subjects' data better than the classic TD model. Similarly, the AC models reflect this same finding, suggesting that this architecture can also account for McDougle et al.'s behavioral results, but only with the presence of gating.

Dual gating is the most feasible AC gating model

All three of the AC gating models are capable of producing both risk averse and risk seeking behavior. However, the actor gating model can only produce risk aversion and risk seeking in tiny fractions of parameter space (Fig. 5E). Similarly, the critic gating model can only produce risk seeking in a tiny fraction of parameter space, though it can produce risk aversion in the majority of its parameter space (Fig. 5F). A large fraction of the dual gating parameter space can produce risk aversion, and a relatively large fraction can produce risk seeking compared to the other two gating models (Fig. 5D).

Although fractions of both the TD gating and AC dual gating model parameter spaces can produce risk seeking, they are much smaller in magnitude than the risk averse and risk neutral fractions. This is due to the fact that a significant amount of gating is required to produce risk seeking behavior. While risk seeking behavior is prevalent in regions with sufficiently low gating rates, its overall volume appears minor because the rest of the space is dominated by risk averse and risk neutral behavior. The fractions of risk seeking behavior in the actor and critic gating AC models are so tiny that it suggests these behaviors are not truly characteristic of these models.

In order to confirm that the larger fraction of volume for risk seeking in the dual gating model truly implies a more feasible gating mechanism, we compared model fits to subject data using the BIC (Fig. 6). All AC models have comparable BIC values for the standard condition. However, the dual gating model has a noticeably lower BIC value for the spatial and binary conditions. Thus, the dual gating model is the only AC gating model that provides a substantial improvement in fit for the spatial and binary conditions compared to the classic AC model.

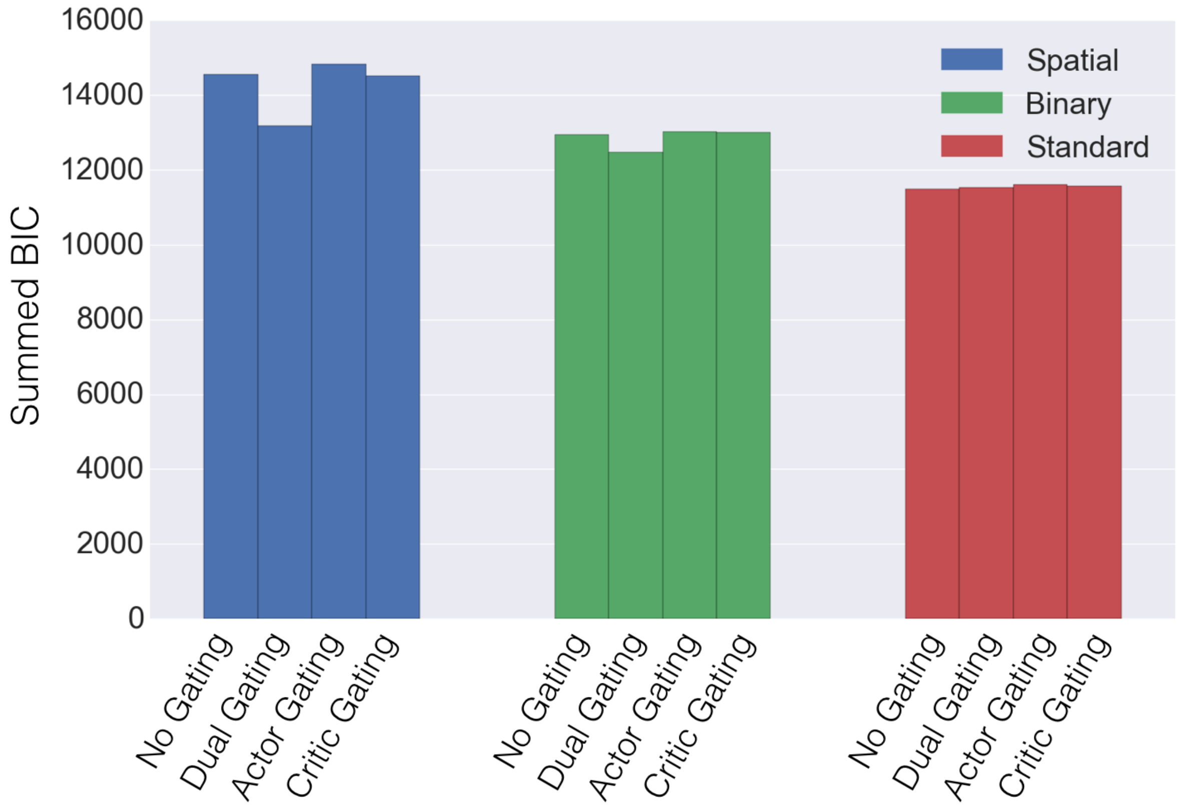


Figure 6: *AC Gating Model Comparison*. Summed BIC values for each model for each experimental condition in McDougle et al. Lower BIC values mean the model better accounts for subject data. While the classic, actor gating, and critic gating models have similar BIC's across conditions, the dual gating model has a noticeable reduction in the spatial and binary conditions, implying a better fit to McDougle et al.'s subject data.

AC models do not provide a better fit to subject data than TD models

It is unclear if the additional flexibility provided by the extra learning rate in the AC models allows for better fits to subject data compared to the TD models. To compare these models in a more direct manner, we again use the BIC (Fig. 7). The AC models do not produce lower BIC values for any of the conditions and instead show slight increases. This suggests that the extra flexibility afforded to the AC models does not provide any behavioral advantage with the current data sets.

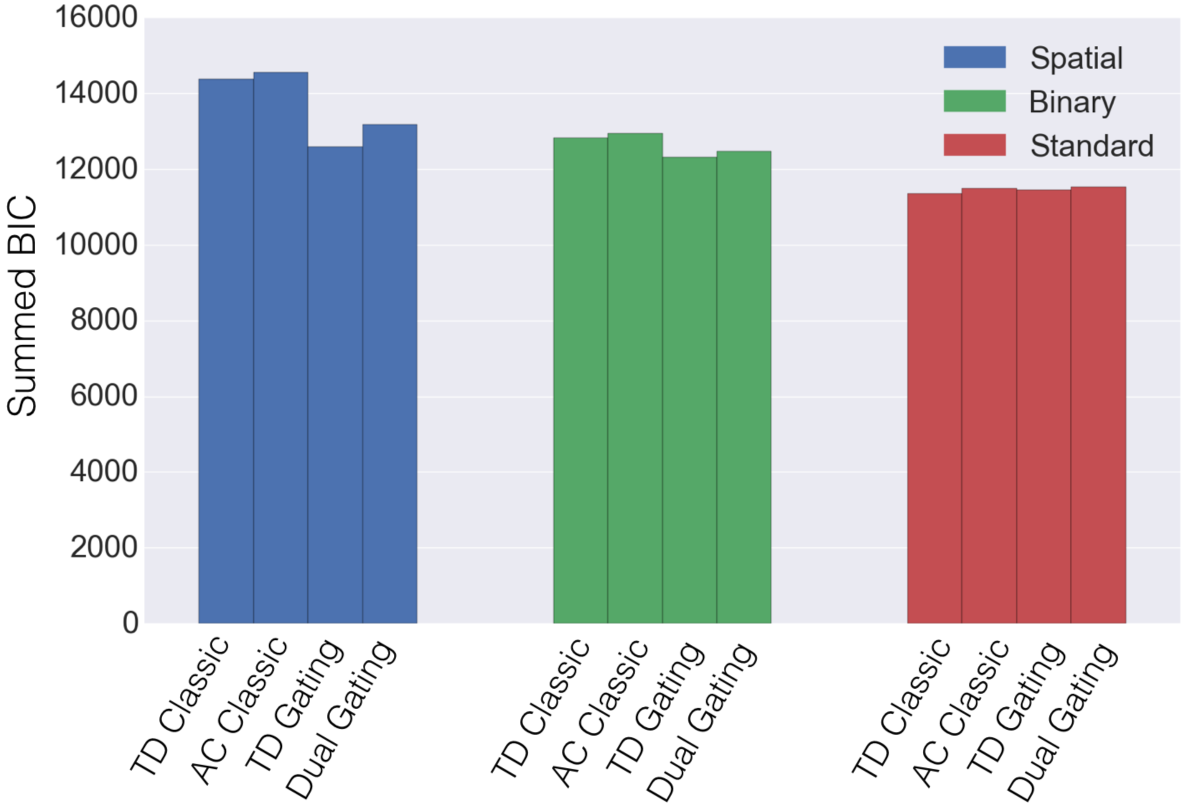


Figure 7: *AC & TD Model Comparison*. Summed BIC values for each model for each experimental condition in McDougle et al. Lower BIC values mean the model better accounts for subject data. The AC models do not outperform the TD models in any of the conditions.

Other parameters besides gating can produce shifts in risk bias

Although gating is necessary to produce risk seeking biases, variations along the other parameter dimensions can also shift risk biases. Moving from high to low values of either α or β can eliminate risk aversion in the TD classic model (Fig. 8A). Low values of β lead to less exploitation of the higher valued safer option, which reduces the risk averse bias. In agreement with past work, risk aversion requires relatively high values of α (Niv et al., 2002). Similarly, risk aversion in the AC classic model requires relatively high values of β and both the actor and critic learning rates (Fig. 8B).

Visualizing the TD gating model's parameter space confirms that the observed transition from risk averse to risk seeking biases between the standard and spatial conditions is only possible via increased gating (lower α_{gate}, Fig. 8C). However, this transition is eliminated with lower values of β. Thus, variation in the amount of exploration appears to be able to mask the effects of gating on changes in risk bias. Similar investigations of the AC dual gating model parameter space reveal the same relationship with β. The AC dual gating model also cannot transition from risk aversion to risk seeking biases without gating.

However, the observed effect of β may simply be an artifact of initializing the value estimates for the targets to 0. Since the starting estimates are 0, the target that first gives points (more likely to be the safe target) will quickly gain preference over the other target. Thus, lower values of β may be needed to allow for hit trials to be experienced with the risky target to overcome this initial bias. Since human subjects were told that the targets would pay out between 1 and 100 points, a more intuitive initial value estimate for each target might be 50 points. Indeed, changing the initial value estimates for the targets to 50 eliminates β's ability to mask the effect of gating on the transition between risk biases (Fig. 8D).

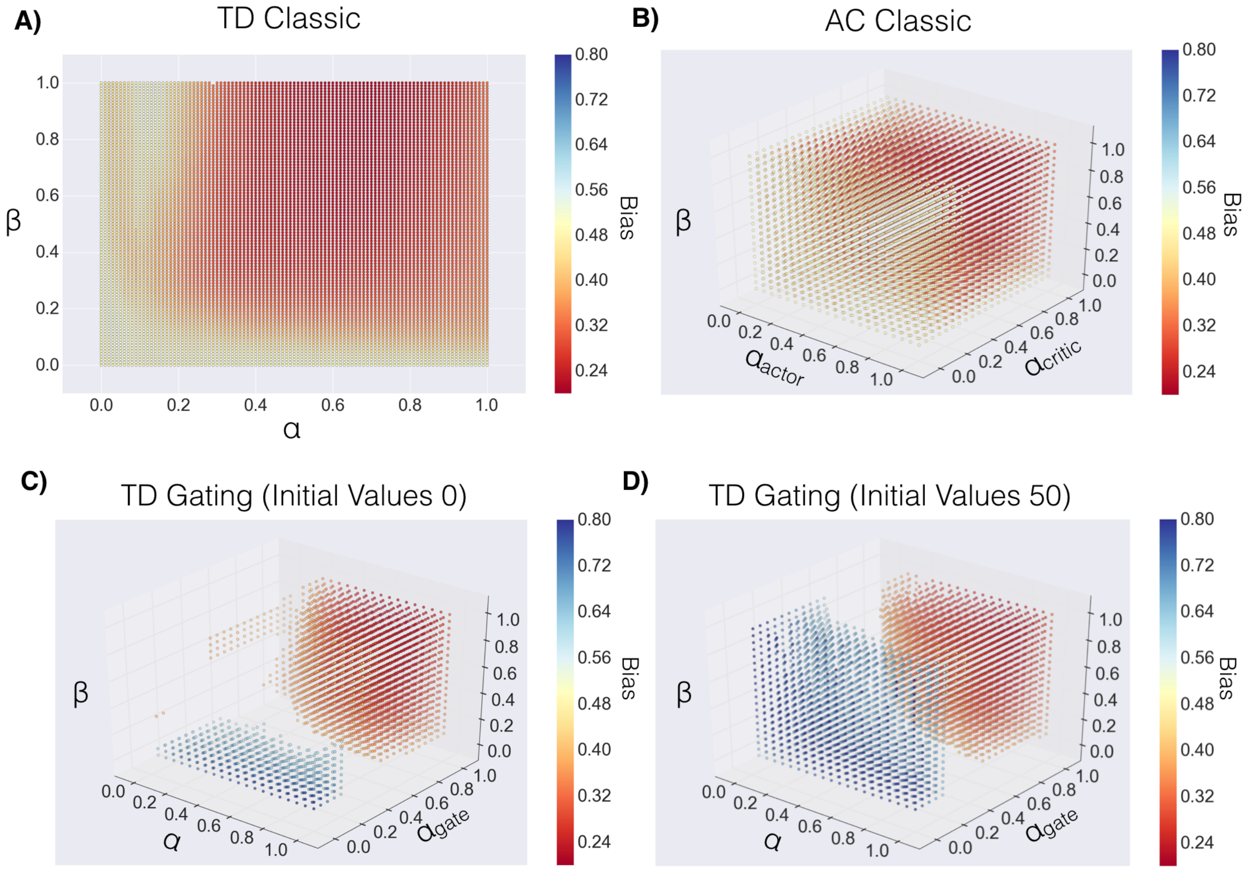


Figure 8: *Model Parameter Spaces*. The color of each point corresponds to the average bias of 100 simulations of the pictured model with parameter settings determined by the axes values. Lower biases (red) correspond to greater risk aversion and higher biases (blue) correspond to greater risk seeking. (**A**) Parameter space for the TD classic model. The parameter space was sampled in 0.01 increments along all parameter dimensions. (**B**) Parameter space for the AC classic model. The parameter space was sampled in 0.05 increments along all parameter dimensions. (**C**) Parameter space for the TD gating model with initial value estimates of 0. The parameter space was sampled in 0.05 increments along all parameter dimensions. Only the risk seeking and risk averse patterns are shown. Uncolored portions of parameter space correspond to risk neutral or perseveration behavior. (**D**) Parameter space for the TD gating model with initial value estimates of 50. The parameter space was sampled in 0.05 increments along all parameter dimensions. Only the risk seeking and risk averse patterns are shown. Uncolored portions of parameter space correspond to risk neutral or perseveration behavior.

Discussion

We found that the AC model can account for subject data, but that gating is necessary to do so. This finding reinforces McDougle et al.'s claim that gating is the underlying factor driving the change in risk attitude between the spatial and standard conditions.

We also showed that the dual gating model is the only AC gating model that can reproduce subjects' risk seeking behavior. The actor gating and critic gating models can both produce risk averse and risk seeking behaviors (in tiny fractions of parameter space), but only the dual gating model shows a better fit to McDougle et al.'s subject data when compared to the AC classic model.

Additionally, we showed that the dual gating model does not better account for subject data than the TD gating model. This means the added flexibility of the separate actor and critic learning rates does not provide any advantage in accounting for subject data. Although the dual gating model does not differentiate itself from the TD gating model, its superiority over the other two AC gating models informs our investigation of possible neural gating substrates.

Finally, we found several other relationships between risk biases and parameters. In line with previous work, we found that low values of learning rates eliminate risk biases. The most surprising relationship found was that risk seeking can only be produced by relatively low values of β. However, further investigation revealed that this may have been a result of initializing the value estimates for the targets to 0. Since the safer target gives points much more frequently than the riskier target, greater exploration may be required to allow for points to be received from the riskier target initially. If this is the case, the relationship should disappear either as the length of the experiment increases or as the initial value estimates increase. We showed that the latter does indeed eliminate this relationship with β.

This raises the question of whether initializing the value estimates to 0 is fully justified. Intuitively, initializing them to 50, the middle of the range provided to subjects, may make more sense. If you knew a target gave between 1 and 100 points, receiving 1 point might seem like an unfavorable outcome, even if you had no other samples from the target yet. Further analyses should consider treating the initial value estimate as a parameter to determine if doing so provides any improvement in accounting for subject data. In any case, raising the value estimates strengthens the case for gating as it eliminates the ability for β to disrupt the shift in risk biases along the gating dimension. Future studies could investigate the effect of prior information about the value of the targets on risk bias and initial choice behavior.

McDougle et al. provide additional evidence that the effect of β on gating when the initial value estimates are 0 could be of interest. In their most recent work, reach feedback on miss trials is constrained to appear in one of two fixed locations, regardless of the subject's reach, instead of varying based on their actual reach as in the original spatial condition. Unpublished work in our lab provides evidence that such feedback still produces SPE's. However, this manipulation fully eliminates the illusion that the subject has any control over the received reward based on the accuracy of their reach. In this new condition, subjects became risk neutral and the full shift to risk seeking behavior was attenuated. Given our findings about β, this effect could still be consistent with the gating hypothesis. The nature of the feedback in this new condition could facilitate more exploitation, which we have shown causes gating to produce risk neutral behavior. Further analysis of this follow-up, in tandem with further analysis of the optimal initial value estimates, should be conducted to determine if observed behavior is well accounted for by both a shift in β and the gating rate.

Biological Plausibility of the Gating Mechanism

Given our findings about the validity of the dual gating AC model, we now turn to a discussion of how these models might relate to neural mechanisms. We use Houk's AC model as a starting point (Houk et al., 1995). In Houk's model, distinct parts of the dorsal striatum serve as the actor and the critic. Learning in both the actor and critic occurs via DA-dependent corticostriatal plasticity mediated by dopaminergic projections from the substantia nigra pars compacta (SNc). Below baseline, DA leads to long term depression (LTD) at the corticostriatal synapse, while above baseline, DA leads to long term potentiation (LTP). Mapping this onto McDougle et al.'s behavioral design implies below baseline DA resulting in corticostriatal LTD on miss trials. Thus, our gating mechanism should prevent this corticostriatal LTD.

The gating hypothesis was primarily motivated by recently discovered disynaptic projections from the deep cerebellar nuclei to the striatum relayed via the thalamus (Hoshi et al., 2005). These projections originate primarily in the dentate, but also in the fastigial and interpositus nuclei to a lesser degree, and terminate in the dorsal striatum. In order for these projections to produce the desired gating effect, they should be able to influence corticostriatal plasticity. Encouragingly, recent physiological study has shown modulation of corticostriatal plasticity by cerebellar input along the cerebellar-basal ganglia projections (Chen et al., 2014). Specifically, they showed that cerebellar depolarization of striatal neurons with simultaneous high frequency cortical input stimulation reversed corticostriatal LTD, which aligns with the desired gating behavior.

It remains unclear whether the deep cerebellar nuclei are sensitive to SPE's. Despite their prominence in many theoretical considerations of the cerebellum, the neural correlates of SPE's have yet to be fully determined. Two pieces of evidence suggest that the cerebellum and more specifically, the deep cerebellar nuclei, may encode representations of such signals. First, Schlerf et al. presented neuroimaging evidence for the presence of SPE correlated activity in the human cerebellum (Schlerf et al., 2012). Second, recent work has presented physiological evidence for the encoding of SPE's in the monkey fasitigial nucleus (Cullen & Brooks, 2015). Thus, it may be feasible that the deep cerebellar nuclei could convey information about SPE's, though no evidence has been given that they are encoded in the dentate, the primary source of the disynaptic projections from the cerebellum to the striatum.

So far we have argued that dual gating in Houk's AC model is consistent with current anatomical and physiological findings. However, the biological feasibility of Houk's AC model and related models have been brought into question (Joel et al., 2002). Specifically, Joel et al. argued that the implementation of the critic by the dorsal striatum is inconsistent with anatomical evidence, and instead proposed that the dorsal striatum serves solely as the actor. The mechanisms discussed up to this point would remain consistent with an actor gating model, but our results suggest an actor only gating model cannot account for McDougle et al.'s behavioral findings. As a result, we now turn to alternative accounts of the biological basis for the critic to determine if they are consistent with potential gating substrates.

In response to criticisms of the original AC design, more recent work has suggested the ventral striatum as an alternative correlate for the critic (O'Doherty, 2004). The ventral striatum has been shown to encode predicted reward (O'Doherty et al., 2002). It also has reciprocal connections with the ventral tegmentum area (VTA), which is primarily composed of dopaminergic neurons that encode RPE's similar to the SNc (Schultz et al., 1997). Finally, the ventral striatum sends projections to the SNc. These three pieces of evidence allow the ventral striatum to form the adaptive critic with the VTA producing the reinforcement signal for the critic and the SNc producing the reinforcement signal for the actor. In order to align with the dual gating model, learning in the critic must also be gated. To our knowledge, there is no clear evidence of cerebellar modulation of either the VTA or ventral striatum, though anatomical projections from the deep cerebellar nuclei to the VTA have been demonstrated in rats (Perciavalle et al., 1989).

We have considered two different subcortical implementations of the AC dual gating model. While Houk's model aligns better with anatomical and physiological characterization of cerebellar-basal ganglia projections, the underlying AC model has faced criticism. Under an alternative subcortical model involving the ventral striatum and VTA as an alternate critic substrate, cerebellar projections to the striatum only align with the actor gating model, and there is minimal demonstrated evidence of cerebellar projections to critic related areas producing physiological effects consistent with gating. Without further studies on relevant anatomy and physiology, our analysis is limited by the lack of biological detail in our AC model. Given the absence of further biological detail in the literature, future models should incorporate more biological detail to allow for the testing of hypotheses about potential gating substrates.

The cerebellum also projects to non-motor regions of cortex (Strick et al., 2009). These projections could serve as an alternate gating pathway for a cortical-based decision-making system, which we have not considered here, in large part because McDougle et al.'s results suggest an implicit, automatic gating mechanism. SPE's have also traditionally been theorized to come from the inferior olive and communicated to the cerebellum via the climber fiber pathways, rather than computed directly in the cerebellum itself (Wolpert et al., 1998). Thus, there may be alternate sources of SPE's or other forms of motor error that may suggest alternate projections subserving the gating mechanism. Future investigations should expand the considered anatomy to look for alternate sources of motor error and gating targets.

Conclusions

AC models are capable of accounting for McDougle et al.'s behavioral results, but gating is necessary to do so. Simultaneous gating of the actor and critic is the only AC gating model that can produce risk seeking behavior, but this dual gating model does not provide a better fit to subject data than the TD gating model. Our consideration of the alignment between biologically proposed AC models and potential gating substrates reveals that the proposed cerebellar-basal ganglia gating pathway is only consistent with the actor gating model. Either an additional critic gating pathway or an alternative mechanism must be considered. Additional behavioral, anatomical, and modeling analyses are needed to determine the full feasibility and nature of a neural gating mechanism.

Acknowledgements

Special thanks to Dr. Matthew Crossley and Professor Richard Ivry for all of their time and guidance over the past two years. Thanks to the Summer Undergraduate Research Fellowship program and staff for their financial and intellectual support of the presented work. Thanks to Professor Ming Hsu for serving as the second reader and to Sam McDougle for providing the model fitting code. Finally, thanks to Darius Parvin for the insight about the relationship between β and the initial value estimates.

References

Barto, A. G. (1995). Adaptive critic and the basal ganglia. In J. C. Houk, J. L. Davis, and D. G. Beiser, (Eds.), *Models of information processing in the basal ganglia* (Cambridge: MIT Press), 215-232.

Bostan, A. C., Dum, R. P., & Strick, P. L. (2010). The basal ganglia communicate with the cerebellum. *Proceedings of the National Academy of Sciences USA*, 107, 8452-8456.

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *Journal of Neuroscience,* 19, 10502-10511.

Calabresi, P., Gubellini, P., Centonze, D., Picconi, B., Bernardi, G., Chergui, K., Svenningsson, P., Fienberg, A. A., & Greengard, P. (2000). Dopamine and cAMP-regulated phosphoprotein 32 kDa controls both striatal long-term depression and long-term potentiation, opposing forms of synaptic plasticity. *Journal of Neuroscience,* 20, 8443-8451.

Chen, C. H., Fremont, R., Arteaga-Bracho, E. E., & Khodakhah, K. (2014). Short latency cerebellar modulation of the basal ganglia. *Nature Neuroscience*, 17, 1767-1775.

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

Cullen, K., & Brooks, J. (2015). Neural correlates of sensory prediction errors in monkeys: Evidence for internal models of voluntary self-motion in the cerebellum. *Cerebellum*, 14, 31-34.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature,* 441, 876-879.

Denrell, J. (2007). Adaptive Learning and Risk Taking. *Psychological Review,* 114, 177-187.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks,* 12, 961-974.

Hoshi, E., Tremblay, L., Féger, J., Carras, P. L., & Strick, P. L. (2005). The cerebellum communicates with the basal ganglia. *Nature Neuroscience*, 8, 1491-1493.

Houk, J. (2005). Agents of the mind. *Biological Cybernetics,* 92, 427-437.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use reward signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (Cambridge: MIT Press), 249–270.

Houk, J., & Wise, S. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex,* 2, 95-110.

Joel, D., Niv, Y., & Ruppin, E. Actor–critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks,* 15, 535-547.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica,* 47, 263-291.

March, J. G. (1996). Learning to be risk averse. *Psychological Review,* 103*,* 309-319.

McDougle, S., Crossley, M., Boggess, M., Ivry, R., & Taylor, J. (in review). Cerebellar dependent error signals are exploited to modulate reinforcement learning.

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

Niv, Y., Joel, D., Meilijson, I., & Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: A simple explanation for complex foraging behaviors. *Adaptive Behavior,* 10, 5-24.

O'Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, 14, 769-776.

O'Doherty, J.P., Deichmann, R., Critchley, H.D., Dolan, R.J. (2002). Neural responses during anticipation of a primary taste reward. *Neuron,* 33, 815-826.

Perciavalle, V., Berretta, S., Raffaele, R. (1989). Projections from the intracerebellar nuclei to the ventral midbrain tegmentum in the rat. *Neuroscience,* 29, 109-119.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review,* 113, 57-83.

Pitt, M. A., Myung, J. I., Montenegro, M., & Pooley, J. (2008). Measuring model flexibility with parameter space partitioning: an introduction and application example. *Cognitive Science*, 32, 1285-303.

Schlerf, J., Ivry, R. B., & Diedrichsen, J. (2012). Encoding of sensory prediction errors in the human cerebellum. Journal of Neuroscience, 32, 4913-4922.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275, 1593-1599.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, 121, 350–354.

Suri, R. E., Bargas, J., & Arbib, M. A. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, 103, 65-85.

Sutton, R.S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. (Cambridge: MIT Press).

Wickens, J. R., Begg, A. J., & Arbuthnott, G. W. (1996). Dopamine reverses the depression of rat corticostriatal synapses which normally follows high-frequency stimulation of cortex in vitro. *Neuroscience*, 70, 1-5.

Wolpert, D. M., Miall, R. C., & Kawato, M. (1998). Internal models in the cerebellum. *Trends in Cognitive Science,* 2, 338-347.