Basal Ganglia-Inspired Functional Constraints Improve the Robustness of Q-value Estimates in Model-Free Reinforcement Learning

Patrick J. Rice (pjrice@uw.edu)
Department of Psychology, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Andrea Stocco (stocco@uw.edu)
Department of Psychology, University of Washington
Campus Box 351525, Seattle, WA 98195 USA

Abstract

Due to the correspondence between the striatal dopamine signal and prediction error signal utilized by model-free reinforcement learning methods, computational psychological research has found much success in modeling the basal ganglia as a biological implementation of a reinforcement learning mechanism. A large majority of these modeling efforts have focused on applying the tenets of reinforcement learning to the proposed functions of the basal ganglia, but few (if any) have attempted to apply crucial aspects of basal ganglia neurophysiology to reinforcement learning mechanisms. Here, we propose a basal ganglia-plausible model that explicitly utilizes two symmetric sets of actions (analogous to the basal ganglias direct and indirect pathways), to simultaneously update value estimates of both available actions (i.e. chosen and not chosen) in the Probabilistic Stimulus Selection (PSS) task. We demonstrate that this proposed model architecture outperforms a standard reinforcement learning model of the PSS task by eliminating the standard model's bias towards estimation of the most valuable available actions, while granting improved resistance to noise in the internal selection process.

apostrophe

Keywords: Reinforcement learning; basal ganglia; dopamine; computational models

Introduction

Model-free reinforcement learning (RL) is a powerful approach for obtaining an optimal long-term action policy in the absence of transition probability and reinforcement functions. In other words, a model-free RL agent must interact with an unknown environment (i.e., sample the environment repeatedly through action) in order to construct an optimal control policy, based on the pattern of reward received by interaction with the environment. This framing of RL methods makes clear their power in modeling human and animal decision-making. Policies refined through RL mechanisms are oriented such that the agent's (i.e., human/animal) actions consider both immediate and future reward, optimized to maximize some value over time. The key idea that enables an agent to determine an optimal policy within an unknown environment is that of temporal-difference (TD) learning (Sutton, 1988).

The ideas behind TD methods have since been expanded, including a proposal by Watkins and Dayan (1992) that defined a TD control algorithm now known as Q-learning. Q-learning is an off-policy method that allows the agent to choose to take non-optimal actions while still estimating an optimal value function. By updating action values based on the best action available while allowing the agent to make inferior choices, this procedure increases the rate of learning under a suboptimal action selection process. Both TD learning and Q-learning have been shown to converge to the optimal value function with probability $P = 1$ (Sutton & Barto, 1998).

However, in some circumstances, these model-free RL methods produce suboptimal results. As defined, these methods emphasize learning of rewarding actions – updating the value of a state/state action (SA) pair increases the likelihood that the agent will choose the action that leads to that state/SA pair when it is next given the opportunity to do so. As a result, even though it converges to an optimal value function, an agent still does not have complete knowledge of its environment – namely, it does not know much (if anything) about the least rewarding states/SA pairs. This is an instance of the general exploration-exploitation trade-off that many models encounter. Lacking knowledge of the least rewarding alternatives is not an issue while the agent has full access to its actions, but what if learned "good" (i.e. optimal) states/SA pairs are blocked from the agent? In this case, the agent cannot take the actions it usually would by following its policy and value function, and as a consequence, cannot act optimally within the "new" environment. In essence, the agent has no knowledge of how to navigate bad options – how to choose the "least bad", when forced to.

The Probabilistic Stimulus Selection Task

A situation in which this circumstance arises is when modeling a well-known psychology task paradigm, the Probabilistic Stimulus Selection (PSS) task (Frank, Seeberger, & O'reilly, 2004). The PSS task is a repetitive, two-alternative forced-choice task made up of two consecutive phases, a *training phase* in which a participant repeatedly makes choices between fixed pairs of stimuli, and a *test phase* where the participant is presented with new combinations of options (see Figure 1).

capitalize

Across both phases, there are six possible stimuli, implemented as symbols that are difficult to describe (in order to make memorization of each stimulus history of success more difficult). Each stimulus carries an intrinsic probability of success, ranging linearly from 20% to 80%. During the training phase, the stimuli are presented a fixed pairs, for a total of three sets: (A,B) (C,D), and (E,F), with associated reward

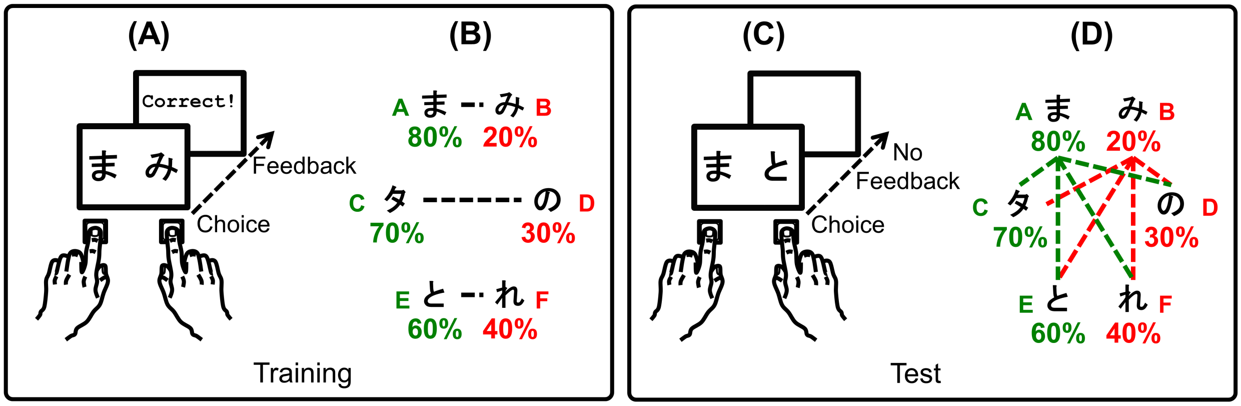


Figure 1: An overview of the Probabilistic Stimulus Selection task. In the *training phase*, participants learn to identify the best option within three pairs. In the]*test phase*, the six options appear in new paired combinations.

AND

probabilities of (80%, 20%), (70%, 30%), and (60%, 40%), respectively. Participants receive feedback regarding the outcome of their decision directly after making a selection. Participants are instructed to attempt to maximize their success by choosing what they believe to be the “correct” option on each trial. Once a participants performance reached a predefined criterion (different for each pair: 65%, 60%, and 50% probability of choosing the higher valued option for the sets of (A,B) (C,D), and (E,F), respectively), the test phase begins. During the test phase, participants are shown all possible combinations of the six stimuli(fifteen total, four times each, for a total of 60 trials), and do not receive feedback upon selection. From the test phase, two different measures are calculated: the participant’s *Choose accuracy*, (the probability of choosing the highest valued alternative (A; 80% reward probability)) of any given fixed set from the training session, when it is paired with any other alternative), and the participant’s *Avoid accuracy* (the probability of not choosing the lowest valued alternative (B, 20% reward probability)) when it is paired with any other alternative (excepting A). These measures can generally be interpreted to be the participant’s tendencies to pursue reward and avoid punishment, respectively.

Remove

Human participants perform close to criterion in the test phase, with an average of about 70% accuracy in both Choose and Avoid (Frank et al., 2004; Frank, Moustafa, Haughey, Curran, & Hutchison, 2007; Stocco et al., 2017).

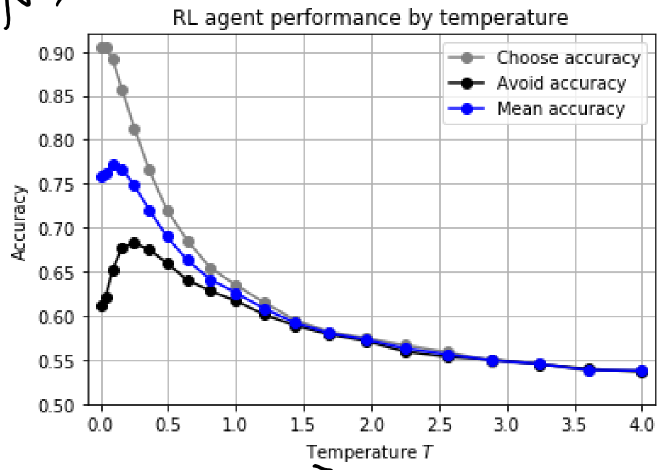
add "accuracies"

Model Comparisons

General Model Implementation

The PSS task poses a number of important constraints for the design of RL agents. In this section, we outline these constraints, and how they were addressed in the implementation of our agents.

The first constrain is that the set of actions available to an agent corresponds to the decision options in the task, that is, the six options $A, B \ldots F$.



Move?

Figure 2: Performance of a canonical RL model in the PSS task for various levels of temperature T

The second constraint is that an agent should be able to generalize the Q value of an action to an different state. This is essential to permit generalization of the Q-values learned during the training phase (Figure 1A and 1B) to the new set of pairs in the test phase (Figure 1C and 1D). A number of mechanisms have been proposed to generalize Q-values to new states. In this paper, we have taken the minimalistic approach of associating all actions to a single state s, but changing the set of actions available at every trial depending on the options presented. Thus, in a trial where the options A and B are presented, only the actions a_A and a_B will be selectable by the agent.

The third constrain is related to the second, and concerns the relationship between subsequent states in the PSS task. Because of the PSS task consists of a sequence of *independent* trials, the probability of a state s_{t+1} following another state s_t does not depend on the action taken a_t. Canonical

RL algorithms based on temporal difference rely on the environment states to be concatenated in some way, since update term for the Q-value of an action taken at state s_t depends on the Q-value of the action actions at state s_{t+1} For example, in the Q-learning algorithm, the error term depends on the *best action* available at state S_{t+1}.

uncap

format the same

$$Q_{s_t,a_t} \leftarrow Q_{s_t,a_t} + \alpha[r_{t+1} + \gamma \max(Q(s_{t+1},a) - Q_{s_t,a_t})] \quad (1)$$

Other algorithms, such as SARSA, similarly rely on the measuring the Q-value of the action taken at state s_t, i.e. $Q(s_{t+1},a_{t+1})$. Since the trials are randomized, however, the contribution of the term $Q_{s_{t+1},a}$ is going to be statistically identical, in the long term, across all states in the long-term. For convenience, in these simulations we set this term to be zero, so that the final learning equation reduces to:

$$Q_{s_t,a_t} \leftarrow Q_{s_t,a_t} + \alpha[r_{t+1} - Q_{s_t,a_t})] \quad (2)$$

Note that, under these conditions, the Q-value of an action a converges to the probability of reward $P(R_t)$ associated with each corresponding option.

Standard RL Model

When the participant in the PSS task is modeled as a Q-learning agent that employs a policy based on the Gibbs function:

$$P(a_i) = \frac{e^{\frac{Q(a_i)}{T}}}{\sum_j e^{\frac{Q(a_j)}{T}}}. \quad (3)$$

is "normalized" the right word?

(under which the probability of the agent choosing a given action increases proportionally with the action's value $Q(s,a)$, normalized by a constant T, defined as the *temperature* of the system. Higher values of T inject more noise into the action selection process, making selection less deterministic), some alarming results are observed. Specifically, the model learns the value of of the desirable options A, C, and E well, reflected as an increasing Choose accuracy as T decreases (Figure 2, grey line). This is the expected behavior of the model – as exploration begins to suggest relatively "better" options, the model quickly switches to exploiting them, learning their true values well in the process [1].

switch to reflect 2?

However, when the Avoid accuracy of the model is inspected, it becomes clear that the model has learned the value of some, but not all, options well. As the value of T begins decreasing, the Avoid accuracy of the model does begin increasing, as the Choose accuracy did. However, the model's Avoid accuracy actually begins decreasing (Figure 2, black line) as T continues to decrease. This indicates that for lower values of T, the model does not sufficiently explore the "bad" options (B, D, and F) during the training phase, and as a consequence, does not value them appropriately. For higher values of T, the model does explore both bad and good options approximately equally – however, it does not value neither good nor bad options appropriately. Additionally, the maximum Avoid accuracy achieved at the point of inflection (less than 70%) is much lower than the maximum Choose accuracy achieved by the model (which is when the value of T is at a minimum; approximately 90%), as well as the Choose accuracy at the point of inflection.

This pattern of Choose and Avoid accuracies over the range of T values tested suggests the existence of an accuracy/bias trade-off – to become more accurate on average for a given option, the model must bias it's action choices to exploiting that option (in other words, the model increases the quality of it's estimates of the "good" options, while becoming more uncertain about the value of the "bad" options). This effect can be seen as tendency of RL agents to converge towards overly optimistic estimates, which has been noted in the literature (Hasselt, 2010). Note that this trade-off effect does not manifest in human performance. To visualize the model's trade-off issues, the model's estimate error (defined as the bias towards choosing a given option, with respect to the probability of avoiding the same option) can be plotted as a function of its mean accuracy (Figure 3). An ideal PSS task agent would be able to obtain unbiased estimates for every level of accuracy (the vertical dashed black line). However, as made clear by Figure 3, the model's estimate error increases as mean accuracy increases–the model becomes more uncertain about its "bad" options in order to do well when presented with "good" options.

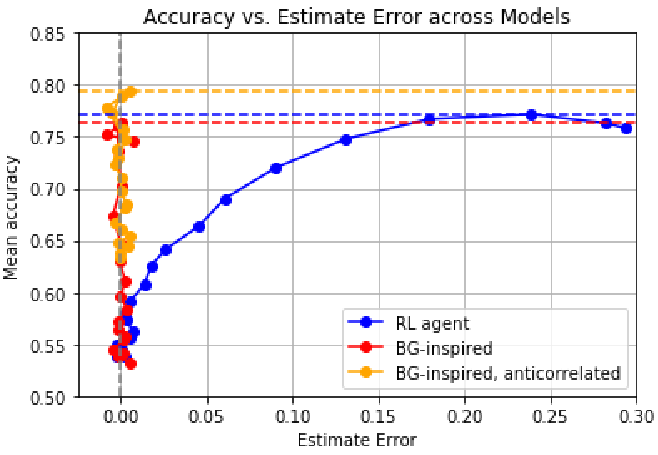


Figure 3: Mean accuracy vs. Q-value estimate errors for the three models examined in this paper

Another way in which this apparent accuracy/bias trade-off can be demonstrated is by defining the model so that it learns the value of NOT choosing actions, rather than the value of choosing actions. In other words, the model chooses to "not choose" a given option, learning the value of such in the pro-

[1] Although here we report the results obtained using Gibb's distribution, the same results have been replicated with another common policy that balances exploration and exploitation, the ε-greedy policy

period

remove

cess. In this case, as the value of T decreases, Avoid accuracy increases while Choose accuracy exhibits the inflection behavior seen in Avoid accuracy under the original model (Figure 2?). Now, the model has learned how to navigate amongst "bad" options–it knows the value of not choosing a given option, and so it "doesnt choose" the "bad" options more often as T decreases. However, it does not learn about the value of "good" options during the learning process.

Basal Ganglia-Inspired RL Method

Reinforcement learning is known to be a reliable method of modeling the function of the basal ganglia (BG) system, a network of subcortical nuclei including the striatum, globus pallidus, substania nigra, and subthalamic nucleus (Alexander & Crutcher, 1990).

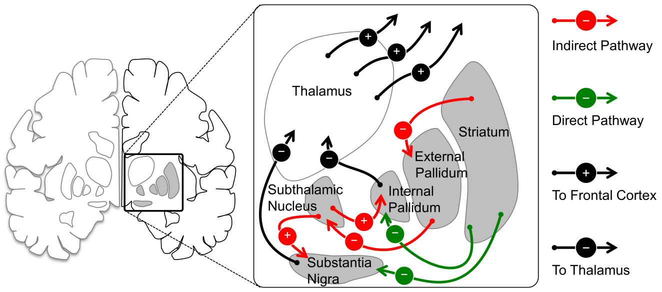


Figure 4: Overview of functional anatomy of the basal ganglia. The main basal ganglia nuclei are in grey; the arrows indicate the major projections between nuclei.The indirect pathways is shown in red, while the direct pathway is shown in green.)

The striatum receives input from cortical structures, and subsequently propagates the signal to later nuclei of the BG through two distinct pathways, termed the "direct" and "indirect" pathways (Smith, Beyan, Shrink, & Bolam, 1998). Of particular interest to neurological/psychological research is the fact that the striatum also receives strong dopaminergic (dopamine; DA) input from the substantia nigra pars compacta (SNc). Dopaminergic signaling originating from the SNc has long been thought to reflect a neural "reward" signal associated with internally-generated action and external stimuli that the organism has learned is (or expects to be) rewarding in some manner, and corresponds closely with the prediction error signal utilized in RL methods (Schultz, 2000; Schultz, Dayan, & Montague, 1997). Additionally, dopaminergic input is a defining characteristic of the "direct" and "indirect" pathways mentioned above – striatal neurons that express D1 receptors (for which DA is an excitatory ligand) form the origin of the direct pathway, while those that express D2 receptors (for which DA is an inhibitory ligand) form the origin of the indirect pathway.

italics?

For the PSS task, although the standard RL model does fairly well overall (approximately 77%), it's performance does not match that of human participants, especially when considering Avoid accuracy. As the model's results demonstrate, it learns well about one set of options (either the "good" options or the "bad" options, depending on if it is learning what to choose or what to not choose, respectively), but it does not do well at valuing all options appropriately at all values of T. Ideally, the model could instead learn the values of choosing an option and not choosing the alternative simultaneously, allowing it to train once in order to appropriately value all possible options. Superficially, there seems to be an obvious compatibility between the necessity for a RL model to simultaneously estimate the value both the "chosen" and "not chosen" alternatives within a PSS trial, and dopamine's opposing influence on the direct and indirect pathways. Would a model-free RL agent with two "action pathways" perform any better than the standard RL model described above?

In order to implement the two-pathway concept, the Q-learning agent described above was modified to include an opposite set of "don't" actions $(\neg A, \neg B, \ldots, \neg F)$, which, when chosen by the agent, result in the selection of the other option that they are paired with. Thus, this agent contains a double set of actions and a stores a double set of Q-values; in this, it is reminiscent of double Q-learning (Hasselt, 2010; Van Hasselt, Guez, & Silver, 2016), an algorithm devised to address the overly-optimistic estimates of the original Q-learning algorithm (Watkins & Dayan, 1992).

The original set of actions $(A, B, \ldots, F)$ can be conceptualized as the set of actions available to be suggested by the direct pathway (restricted by actions possible within the current state), while the "antiset" can be conceptualized as the set of actions available to be suggested by the indirect pathway (also restricted by the state). So, if the current trial allows for actions A and B, and the agent selects the indirect pathway's action $\neg A$, the result is the selection of option B. However, if the current trial allows for actions A and C, and the agent selects $\neg A$, the result is the selection of option C.

change action refs to format above

Figure 5 shows that the simple addition of an "indirect pathway" to the RL model results in a marked absence of the bias observed in the standard RL model–as the value of T decreases, both choose and avoid accuracies increase commensurately. As such, the model no longer needs to "trade-off" increasing the accuracy for one class of action by becoming less confident in the valuations of the other class of action. Instead, for every choice made, it simultaneously learns both the value of the option chosen, and the value of not choosing the alternative. However, note that the maximum Choose and Avoid accuracies of the BG-plausible model do not quite achieve the same level of accuracy as the standard RL models–the uncertainty that the standard model had been attributing to the option not chosen has now been distributed across both available options. Figure 3 demonstrates that overall, the BG-plausible model (red line) achieves essentially the same level of global mean accuracy as the standard RL model (blue line), but *without* the cost of increasing estimate error.

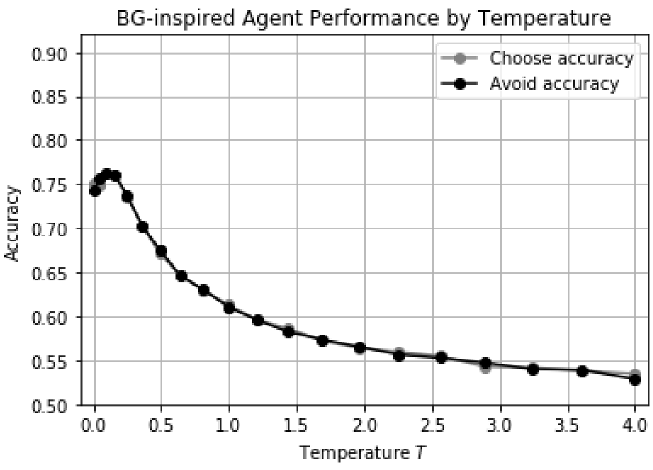


Figure 5: Performance of the BG-inspired reinforcement learning agent in the PSS task for various levels of temperature T. Note that there is no difference in the Choose and Avoid accuracies

Making the Model More Plausible

As described, this implementation of "direct" and "indirect" pathways in the RL model does well at capturing the competition between the direct and indirect pathways of the BG, and alleviates the problem of increasing estimate error with increasing accuracy. However, the BG-plausible model still performs similarly to the standard RL model in terms of global mean accuracy, indicating that although the BG-plausible model has improved ability to estimate the value of all options in the environment, this does not translate to improved fitness within the environment. However, just as the standard models were missing a crucial aspect of BG physiology (the presence of dual pathways), the BG-plausible model is missing a crucial feature of these dual pathways – the fact that DA signaling has opposite effects on the direct (excitation, mediated through D1 receptors) and indirect (inhibition, mediated through D2 receptors) pathways.

To capture this aspect of BG neurodynamics, the BG-plausible RL model was modified so that the learning algorithm results in *opposite* changes for the actions to the two pathways (an anti-correlated BG-plausible model). Specifically, if action A was selected and resulted in an update of it Q-value of size δ, then the Q value of the corresponding anti-action $\neg A$ would be updated by the quantity $-\delta$. As in the biological basal ganglia, this mechanisms forces the values of one set of actions to be anti-correlated to the values of the other set. Note that, importantly.

Figure 6 shows the results of simulations ran with this model. At minimum values of T, the maximum mean Choose and Avoid accuracies increase slightly (when compared to the original BG-plausible model). Figure 3 shows that, similar to the original BG-plausible model (red line), the mean accuracy of the anti-correlated BG-plausible model (yellow line) increases without a subsequent increase in estimate error. Additionally, the small increase in Choose and Avoid accuracies at minimum values of T translate into significantly better overall performance for the anti-correlated BG-plausible model.

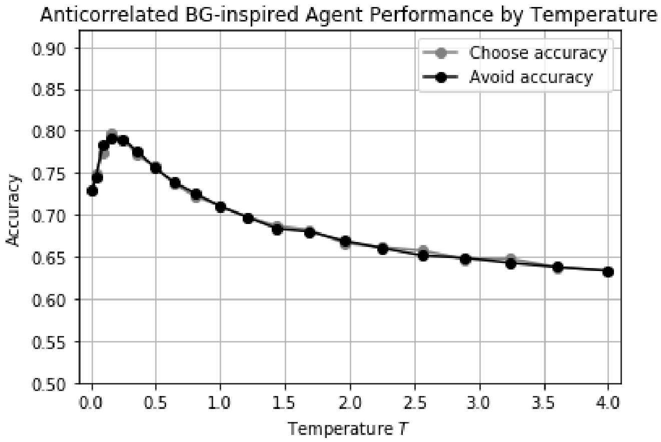


Figure 6: Performance of the anticorrelated, BG-inspired RL-learning model in the PSS task for various levels of temperature T

However, what is most striking about the anti-correlated BG-plausible model is that at relatively large values of T (where the action selection process is noisy), the model performs much better than either the original BG-plausible model, or the standard RL models. This is an indication that the presence of the anti-correlated pathways in the second BG-plausible model bestow a greater resistance to internal noise than the original BG-plausible model and standard RL models possess. Figure 7 more clearly demonstrates this effect: across the range of tested values of T, the mean accuracies of the original BG-plausible model are almost identical to the standard RL model. However, across the same range of T values, the anti-correlated BG-plausible model performs much better in almost every circumstance.

The model does not perform as well as the standard "original BG" models only when the value of T is very close to zero, indicating almost no noise in the action selection process (an unrealistic assumption for biological systems). A similar analysis can be performed for the model's estimate error, as seen in Figure 3. This again shows that for every tested value of T, there is little or no difference between either BG-plausible model – the presence of the two pathways allows each model to accurately estimate the value of both Choose (A, C, and E) and Avoid (B, D, and F) options. However, the standard RL model shows significant estimation biases as the lowest levels of noise, when the model's performance is at a maximum.

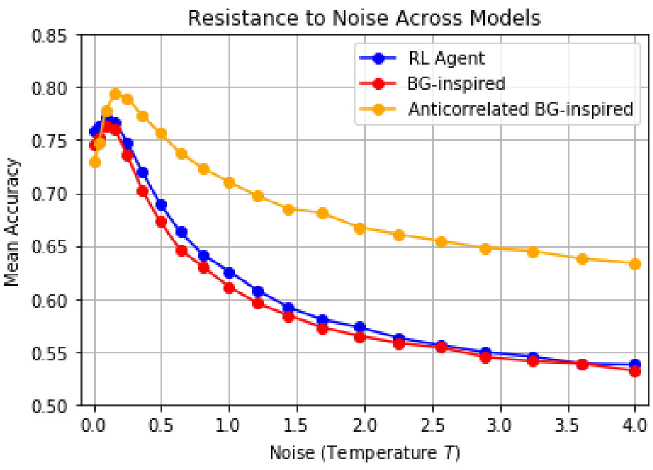


Figure 7: A direct comparison of the mean accuracy of three RL models tested in this paper. *Blue*: Standard RL model; *Red*: BG-inspired model; *Yellow*: Anti-correlated BG-inspired model

.

Conclusions

In conclusion, the improved performance of the BG-plausible RL models implies that psychological researchers looking to model the functions of the basal ganglia could do well by taking inspiration from the characteristics of the phenomena they model, even when the modeling effort is largely theoretical. Addition of the models opposed update pathways, representative of the well-known direct and indirect pathways within the basal ganglia, allowed the original BG-plausible model to properly estimate the value of both the “good” (relatively high probability of reward) and “bad” (relatively low probability of reward) options available in the PSS task, eliminating the bias towards “good” options displayed by the standard RL model. In addition, by forcing the updates of the two pathways to be anti-correlated (thereby mimicking the opposed excitatory/inhibitory effect of dopamine on the direct and indirect pathways), the model displayed a marked resistance to greater levels of noise within the selection mechanism.

Acknowledgments

This research was suported by a grant from the Office of Naval Research (ONRBAA13-003) entitled Training the Mind and Brain: Investigating Individual Differences in the Ability to Learn and Benefit Cognitively from Language Training” and by a start-up grant from the University of Washington to Andrea Stocco. Simulations and code for this study can be found at the Cognition and Cortical Dynamics Laboratory’s GitHib repository: https://github.com/UWCCDL/BGRL.

References

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: Neural substrates of parallel processing. *Trends in neurosciences*, *13*, 266–271.

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*, *104*(41), 16311–16316.

Frank, M. J., Seeberger, L. C., & O’reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(5703), 1940–1943.

Hasselt, H. V. (2010). Double q-learning. In *Advances in neural information processing systems* (pp. 2613–2621).

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience*, *1*, 199–207.

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

Smith, Y., Beyan, M. D., Shrink, E., & Bolam, J. P. (1998). Microcircuitry of the direct and indirect pathways of the basal ganglia. *Neuroscience*, *86*, 353–388.

Stocco, A., Murray, N. L., Yamasaki, B. L., Renno, T. J., Nguyen, J., & Prat, C. S. (2017). Individual differences in the simon effect are underpinned by differences in competitive dynamics of the basal ganglia: An experimental verification and a computational model. *Cognition*, *164*, 31–45.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.

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

Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. In *Aaai* (pp. 2094–2100).

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279–292.