Multiplicity of control in the basal ganglia: Computational roles of striatal subregions

Aaron M Bornstein and **Nathaniel D Daw**
Center for Neural Science and Psychology Department, New York University, 4 Washington Place, New York, NY 10003

Abstract

The Basal Ganglia, in particular the striatum, are central to theories of behavioral control, and often identified as a seat of action selection. Reinforcement Learning (RL) models - which have driven much recent experimental work on this region - cast striatum as a dynamic controller, integrating sensory and motivational information to construct efficient and enriching behavioral policies. Befitting this informationally central role, the BG sit at the nexus of multiple anatomical "loops" of synaptic projections, connecting a wide range of cortical and sub-cortical structures. Numerous pioneering anatomical studies conducted over the past several decades have meticulously catalogued these loops, and labeled them according to the inferred functions of the connected regions. The specific cotermina of the projections are highly localized to several different subregions of the striatum, leading to the suggestion that these subregions perform complementary but distinct functions. However, until recently, the dominant computational framework outlined only a bipartite, dorsal/ventral, division of striatum. We review recent computational and experimental advances that argue for a more finely fractionated delineation. In particular, experimental data provides extensive insight on unique functions subserved by the dorsomedial striatum (DMS). These functions appear to correspond well to theories of a "model-based" RL subunit, and may also shed light on the suborganization of ventral striatum. Finally, we discuss the limitations of these ideas and how they point the way toward future refinements of neurocomputational theories of striatal function, bringing them into contact with other areas of computational theory and other regions of the brain.

Keywords

Reinforcement learning; Basal Ganglia; Action Selection; Model-based learning; Goal-directed; POMDP

Introduction

Perhaps more than any other brain areas, recent advances in understanding of the basal ganglia (BG) have been driven by computational models. This is largely due to the fact that core functions commonly ascribed to the BG — action selection and value learning — have been the subject of intensive study in both economics and computer science, particularly the subfield of artificial intelligence known as reinforcement learning (RL) [1]. Theories from

these areas propose mathematical definitions for quantities relevant to these functions and step-by-step procedures for computing them. Accordingly, these models have rapidly progressed from general frameworks for interpreting data toward playing a more integral quantitative role in experimental design and analysis, and now often serve as explicit hypotheses about trial-by-trial fluctuations in biological signals, such as action potentials or blood oxygenation level dependent (BOLD) signals. The poster children for this approach are the influential, albeit controversial, temporal difference (TD) learning models, which describe a reward prediction error (RPE) signal that has proved a strong match to phasic firing of midbrain dopamine neurons as well as BOLD in the ventral striatum (VS) [2–7]. The present review considers recent work that has expanded upon this initial achievement, shedding further light on the computational and functional suborganization of striatum, and then considers questions raised by this work in light of other empirical data and computational modeling that, together, point the way for future work in this area.

The suborganization of striatum

Although anatomical considerations such as the topographical gradient of afferents from cortex to striatum have long suggested considerable functional heterogeneity (for instance, five distinct corticostriatal loops in one seminal review [8]), there have until recently been surprisingly few clear correlates of this presumptive suborganization in unit recordings or functional neuroimaging within BG. In parallel, there have been few functional subdivisions suggested among the dominant computational models to motivate or guide the search for such prolific variegation.

In RL models, the primary early progress on this question was the functional breakdown between learning to predict rewards (“critic”) and, guided by these predictions, learning advantageous stimulus-action policies (“actor”) [2]. This was suggested [3] to correspond to a two-module breakdown in striatum between motoric actor functions dorsally and evaluative critic functions ventrally, an idea which resonates with data both from rodent lesions [9, 10] and human functional neuroimaging [11–13].

The actor/critic formalizes a classic psychological distinction between Pavlovian learning (about stimulus-outcome relationships), and instrumental learning (about which actions are advantageous). However, psychologists have long known that these functions are further decomposable. Notably, instrumental learning comprises subtypes that guide actions using different learned representations: “habitual” actions based on stimulus-response associations, vs. “goal-directed” actions supported by representation of the particular goal (such as food) expected for an action [14, 15]. This distinction is typically probed using manipulations that alter action-outcome contingency or outcome value, such as reward devaluation, and then evaluating the subsequent effect on responding. After extensive training, behavior can become insensitive to these manipulations, suggesting an isolated reliance on stimulus-response representations, i.e. habits.

A raft of recent theoretical work exploits a parallel distinction in RL models [16–22]. TD theories, such as the actor/critic, correspond well to classical ideas about stimulus-response habits and how they are reinforced [23, 24], but these “model-free” algorithms cannot explain behavioral phenomena associated with goal-directed behavior such as devaluation sensitivity or latent learning [25]. However, an additional family of “model-based” RL algorithms describe how learning about the structure of an environment can be used to evaluate candidate actions online. This evaluation, typically implemented by some sort of simulation, inference or preplay using a *forward model* of the task (analogous to an action-outcome representation or cognitive map) [26–29], can plan new actions or re-evaluate old ones drawing on information other than simple reinforcement history.

Much recent experimental research has been guided by the discovery, via lesions of rodent striatum, that these behaviors are supported by distinct subregions of dorsal striatum, respectively lateral and medial (DLS and DMS) [30–33]. The proposal that a model-free TD actor in DLS is accompanied by an additional model-based RL system in DMS has considerable promise. First, it preserves the substantial successes of the TD/dopamine models while correcting some of their shortcomings: for instance its redundancy of control may help to explain why lesions to dopaminergic nuclei do not prevent all instrumental learning [34]. Conversely, just as TD models have helped to shed light on dopaminergic habit mechanisms, model-based RL may provide a framework for understanding neural mechanisms for goal-directed evaluation.

Such work is at a very early stage; although putative correlates of model-based computations have been reported throughout a very wide network [35–43], probably the most developed data thus far concern spiking correlates for both prospective locations (in hippocampus) and associated rewards (in ventral striatum), suggesting a circuit for model-based evaluation of candidate trajectories [26, 44, 45]. Complementary to such "preplay" phenomena, "replay" of neural sequences may play a similar, but offline, role in updating stored (e.g. model-free) value predictions [46–49], perhaps by sampling model-based trajectories estimated from a cognitive map [50].

Theories of model-based and model-free RL in the basal ganglia envision parallel circuits. This is consistent with findings from lesion studies that the two learning processes appear to evolve side-by-side [30, 31], even though they tend to dominate behavior serially: progressing, with training, from goal-directed to habitual responding. Several features of DMS and DLS unit recordings appear to reflect these differential temporal dynamics, with task-related responsivity in DMS peaking early in training (and in retraining, following task changes) [51, 52]; as well as with changes in ensemble responses [53] and several measures of synaptic potentiation [54] peaking earlier in DMS than DLS.

The last results [54] resonate with at least two other predictions from RL models. The authors measure a greater concentration of dopamine D2 receptors on neurons in DLS (relative to DMS). That these receptors are uniquely sensitive to extrasynaptic tonic dopamine concentrations (which are several orders of magnitude lower than those resulting from phasic bursts) supports a previously posited computational role for tonic dopamine in the modulation and motivational control of habitual expression [18]. Further, the D2-containing neurons (which are known to primarily project into the "indirect", striatopallidal, pathway) were also a primary site of synaptic potentiation during behavioral training. This observation is consistent with a body of computational and experimental work suggesting that these receptors are involved in learning as well as expression, perhaps specifically in learning which actions to avoid [55, 56]

Questions and anomalies

At the same time, many of these studies point to three serious questions for the RL models: on their overall architecture, their mechanisms for learning, and how they are deployed during choice.

Architecture and the model-based critic

First, the basic project of rescuing model-free actor/critic theories of the dopamine system by augmenting them with a separate and parallel model-based RL system is challenged by a number of recent results suggesting that even areas associated with the putatively model-free critic (including ventral stratum, [42, 43, 45], downstream ventral pallidum [39], and RPE units in the dopaminergic midbrain [59]) all show properties such as sensitivity to

devaluation that are indicative of model-based RL and not easily explained by the standard model-free TD theories (see also [60]). These results suggest that the two hypothetical systems are either more interacting or hybrid than separate, consistent with the overlapping "loop" architecture suggested by anatomical studies [8, 61]. An intriguing alternate suggestion is that the ventral striatal critic also consists of dissociable subcomponents for model-based and model-free Pavlovian evaluation (Figure 1). Indeed, psychologists distinguish preparatory and consummatory forms of Pavlovian conditioning, which may involve distinct circuits in the core and shell of ventral striatum [33, 62–64]. This distinction again closely tracks that between model-based and model-free RL, in that consummatory Pavlovian responses reflect knowledge of the particular outcome expected (suggesting they are derived from predictions using a world model), whereas preparatory responses, like a model-free critic's predictions, are not outcome-specific [65].

Learning and hierarchical RL

A related issue is that the considerable algorithmic differences between model-based and model-free RL approaches seem poorly matched to the basic isomorphism of circuitry between different parts of the BG [8]. While the model-free actor and critic both learn from the same error signal operating on different inputs — thought to be consistent with a similar dopaminergic input driving plasticity in both ventral and dorsolateral striatum — the representations used for model-based RL seem to require quite different teaching signals and learning rules [41], offering no obvious role, in these theories, for a dopaminergic RPE in DMS.

One possible direction for resolving this question arises from a somewhat different empirical and theoretical take on the function of DLS, in the "chunking" of behavioral sequences. DLS neurons (and, importantly, not those in DMS) tend with training to cluster their responsivity at the beginning and end of a trial [53], and lesions of DLS [66] (and of a prefrontal area that may be afferent to it [32]) also suggest a causal role in behavioral chunking. In RL, such chunking of actions into multiaction "options" is formalized by a family of "hierarchical" RL models [67]. Taken as an organizing principle for striatum (e.g., with policies operating on elemental actions represented in DMS and, moving laterally, policies on progressively more chunked options), this model has the appealing feature that all levels of the hierarchy learn from a similar (in this case, model-free) RPE signal and a common learning rule.

However, although there appears to be some informal resonance between a stimulus-response habit and an automatized behavioral sequence, in RL, the inclusion of options generally crosscuts the distinction between model-based and model-free evaluative strategies. Since the latter distinction has been used to explain the signature phenomena (such as devaluation sensitivity) tying DMS and DLS to goal-directed and habitual instrumental behaviors, additional theoretical work will be needed to understand if these these two approaches can be blended, e.g. using model-based hierarchical RL, to formalize both chunking and devaluation phenomena together.

Arbitration and choice under uncertainty

A third major question raised by theories involving multiple, parallel reinforcement learners is how the brain arbitrates between the two systems' choices. One theoretical proposal is that the predictions of model-free and model-based reinforcement learners may be competitively combined based on the uncertainty about their predictions [17]. In population code representations, uncertainty may be carried by the entropy of neuronal firing across the population [68, 69]. Indeed, over training Thorn et al. [53] observed differential modulations

in population entropy in DLS and DMS, with the DMS representation becoming structured more quickly but ultimately overtaken in this measure by DLS.

Nevertheless, this theory has little to say about the more dynamic processes or mechanisms by which the brain combines these uncertain estimates. The accumulation of multiple noisy evidence sources has, however, been studied extensively in another heretofore largely distinct area of theoretical and experimental work on decision making about noisy sensory displays. Here, reaction times, errors, and ramping activity of neurons in posterior parietal cortex are famously captured by Bayesian models of the accumulation of evidence about stimulus identity [68, 70]. By these models, sensory decision regions interpret stimuli by drawing sucessive samples of noisy sensory input as represented in upstream sensory cortices (which may themselves incorporate a sensory prior that develops to match the distribution of natural stimuli [71]). This work is rapidly coming directly into contact with research on RL and the BG for a number of reasons.

For one, the success of these Bayesian sequential sampling models is not limited to purely sensory tasks involving the analysis of noisy percepts. Notably, they also capture human behavior in tasks involving more affective, value-driven choices, such as pricing or choosing between snack foods [20, 72, 73]. Thus, goal-directed valuation, too, may involve accumulating stochastic samples, here presumably drawn from memory rather than a noisy percept [74–77]. This implies a rather different mechanism for model-based evaluation than the more systematic tree search or Bayesian graph inference so far hypothesized [17, 28], though perhaps one not incompatible with the relatively noisy preplay phenomena observed neurally [45, 50, 78]. Such a procedure for computing model-based values could incorporate uncertainty-weighted habit information (e.g. as a prior), suggesting a dynamic solution to the arbitration problem. In all these respects, it is interesting that in a sensory decision task, primate caudate neurons display activity related to evidence accumulation not unlike that seen in parietal cortex [79].

These models, finally, speak to the BG's connections with a broader anatomical and computational universe. Typical sensory and RL decision tasks exercise almost entirely complementary functions: in one case, analyzing a noisy sensory stimulus with the response rule well defined, and, in the other, figuring out which candidate response is most valuable with no perceptual uncertainty. There has been considerable interest in how the neurocomputational mechanisms that have been characterized for each function, separately, might interact in more complicated tasks involving both value learning and perceptual (or, in RL terms, "state") uncertainty [80–83]. Models based on RL for so-called partially observable Markov decision processes (POMDP) [80, 82, 84] suggest that these two mechanisms can operate serially: a cortical perceptual inference module infers a distribution over possible stimuli and this serves as input for a BG RL module operating much as before. Such a model explains dopaminergic responses in a perceptual inference task as related to prediction error as the cortical model "figures out" whether the animal is facing an easy (likely rewarded) or hard trial [82, 85]. This approach may also provide a route toward explaining dopaminergic responses related to seeking information about future reward prospects [86].

Last, since the perceptual system, on this view, must in general base its percepts on learning about the statistical structure of the task and percepts, this idea situates the RL circuit alongside substantial recent work on Bayesian models of learning latent structure [87–90]. More generally, such structure learning is a valuable component of an efficient model-based system. Indeed, in realistic RL tasks involving perceptual uncertainty, such latent structure learning is also an important component of learning the world model for model-based RL. Thus, exploring how inference guides the construction and employment of associative

representations may ultimately provide a synthesis between cortical belief computations and model-based striatal RL.

Acknowledgments

The authors are supported by a Scholar Award from the McKnight Foundation, a NARSAD Young Investigator Award, Human Frontiers Science Program Grant RGP0036/2009-C, and NIMH grant 1R01MH087882-01, part of the CRCNS program. We thank Dylan Simon for helpful ideas on heirarchical RL, and Amitai Shenhav and Fenna Krienen for comments on an earlier version of this manuscript.

References and recommended reading

1. Sutton, R.; Barto, A. Reinforcement learning: An introduction. MIT Press; Cambridge, MA: 1998.
2. Barto, AC. Adaptive Critics and the Basal Ganglia. Vol. 1. The MIT Press; 1994. p. 215-32.
3. Montague PR, Dayan P, Sejnowski TJ. A Framework for Mesencephalic Predictive Hebbian Learning. Journal of Neuroscience. 1996;76(5):1936–47.
4. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal Difference Models and Reward-Related Learning in the Human Brain. Neuron. 2003;28:329–37.
5. McClure S, Berns G, Montague P. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 2003;38(2):339–46. [PubMed: 12718866]
6. Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. The Journal of neuroscience. 2009;29(31):9861–74.10.1523/JNEUROSCI.6157-08.2009 [PubMed: 19657038]
7. Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. The Journal of neuroscience. 2009;29(47):14701–12.10.1523/JNEUROSCI.2728-09.2009 [PubMed: 19940165]
8. Alexander GE, Delong MR, Strick PL. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annals of the New York Academy of Sciences. 1986;9:351–81.
9. Packard M, Knowlton B. Learning and memory functions of the basal ganglia. Annu Rev Neurosci. 2002;25:563–93. [PubMed: 12052921]
10. Cardinal R, Parkinson J, Lachenal G, Halkerston K, Rudarakanchana N, Hall J, et al. Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. Behavioral Neuroscience. 2002;116(4):553–67. [PubMed: 12148923]
11. O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable Roles of Ventral and Dorsal Striatum in Instrumental. Science. 2004;304:452–4.10.1126/science.1094285 [PubMed: 15087550]
12. Tricomi E, Delgado MR, Fiez JA. Modulation of Caudate Activity by Action Contingency. Neuron. 2004;41:281–92. [PubMed: 14741108]
13. Tricomi E, Balleine BW, O'Doherty JP. A specific role for posterior dorsolateral striatum in human habit learning. European Journal of Neuroscience. 2009;29(11):2225–32.10.1111/j.1460-9568.2009.06796.x [PubMed: 19490086]
14. Adams C, Dickinson A. Instrumental responding following reinforcer devaluation. The Quarterly Journal of Experimental Psychology Section B. 1981;33(2):109–21.
15. Adams C. Variations in the sensitivity of instrumental responding to reinforcer devaluation. The Quarterly Journal of Experimental Psychology Section B. 1982;34(2):77–98.
16. Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks. 1999;12(7–8):961–74. [PubMed: 12662639]
17. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nature Neuroscience. 2005;8(12):1704–11. [PubMed: 16286932]
18. Niv Y, Joel D, Dayan P. A normative perspective on motivation. Trends in Cognitive Sciences. 2006;10(8):375–81.10.1016/j.tics.2006.06.010 [PubMed: 16843041]

19. Redish A, Jensen S, Johnson A. Addiction as vulnerabilities in the decision process. Behavioral and Brain Sciences. 2008;31(04):461–87.
20. Rangel A, Camerer C, Montague P. A framework for studying the neurobiology of value-based decision making. Nature Reviews Neuroscience. 2008;9(7):545–56. [PubMed: 18545266]
21. Balleine B, Daw N, O'Doherty J. Multiple Forms of Value Learning and the Function of Dopamine. Neuroeconomics: decision making and the brain. 2008
22. Frank MJ, Doll BB, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience. 2009;12(8):1062–8.10.1038/nn.2342 [PubMed: 19620978]
23. Suri R, Schultz W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience. 1999;91:871–90. [PubMed: 10391468]
24. Maia T. Two-factor theory, the actor/critic model, and conditioned avoidance. Learning and Behavior. 2010;38(1) [PubMed: 20065349]
25. Tolman EC. Cognitive Maps in Rats and Men. Psychological Review. 1948;55(4):189–208. [PubMed: 18870876]
26. Johnson A, Meer MAAVD, Redish AD. Integrating hippocampus and striatum in decision-making. Current Opinion in Neurobiology. 2007:692–7.10.1016/j.conb.2008.01.003 [PubMed: 18313289]
27. Addis D, Pan L, Vu M, Laiser N, Schacter D. Constructive episodic simulation of the future and the past: Distinct subsystems of a core brain network mediate imagining and remembering. Neuropsychologia. 2009;47(11):2222–38. [PubMed: 19041331]
28. Botvinick M, An J. Goal-directed decision making in prefrontal cortex: a computational framework. Advances in Neural Information Processing Systems. 2009:169–76.
29. Fermin A, Yoshida T, Ito M, Yoshimoto J, Doya K. Evidence for Model-Based Action Planning in a Sequential Finger Movement Task. Journal of Motor Behavior. 2010;42(6):371–9. [PubMed: 21184355]
30. Yin HH, Knowlton BJ, Balleine BW. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. European Journal of Neuroscience. 2004;19:181–9.10.1046/j.1460-9568.2003.03095.x [PubMed: 14750976]
31. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. European Journal of Neuroscience. 2005;22(2):513–23.10.1111/j.1460-9568.2005.04218.x [PubMed: 16045504]
32. Balleine BW, Liljeholm M, Ostlund SB. The integrative function of the basal ganglia in instrumental conditioning. Behavioural Brain Research. 2009;199(1):43–52.10.1016/j.bbr.2008.10.034 [PubMed: 19027797]
33. Yin HH, Ostlund SB, Balleine BW. Reward-guided learning beyond dopamine in the nucleus accumbens: The integrative functions of cortico-basal ganglia networks. European Journal of Neuroscience. 2008;28(8):1437–48.10.1111/j.1460-9568.2008.06422.x.Reward-guided [PubMed: 18793321]
34. Berridge K. The debate over dopamine's role in reward: the case for incentive salience. Psychopharmacology. 2007;191(3):391–431. [PubMed: 17072591]
35. Valentin V, Dickinson A, O'Doherty J. Determining the neural substrates of goal-directed learning in the human brain. Journal of Neuroscience. 2007;27(15) [PubMed: 17428979]
36. Frank M, Moustafa A, Haughey H, Curran T, Hutchison K. Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of Sciences. 2007;104(41):16311–6.
37. Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. Journal of Neuroscience. 2006;26(32) [PubMed: 16899731]
38. Hampton A, Bossaerts P, O'Doherty J. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences. 2008;105(18)
39. Tindell A, Smith K, Berridge K, Aldridge J. Dynamic Computation of Incentive Salience: Wanting What Was Never Liked. Journal of Neuroscience. 2009;29(39) [PubMed: 19793980]

40••. den Ouden HEM, Daunizeau J, Roiser J, Friston KJ, Stephan KE. Striatal Prediction Error Modulates Cortical Coupling. Journal of Neuroscience. 2010;30(9):3210–9. Using a novel dynamic causal model (DCM) approach, the researchers demonstrate that striatal prediction errors influence, in a trial-by-trial fashion, the degree of functional coupling between visual and motor regions during a serial response task, with greater prediction error leading to stronger connectivity. Similar to the work of Law and Gold [83], this study supports a causal role for the basal ganglia in establishing and tuning cortico-cortico connectivity, via a prediction error signal. 10.1523/JNEUROSCI.4458-09.2010 [PubMed: 20203180]

41. Gläscher J, Daw ND, Dayan P, O'Doherty JP. States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron. 2010;66(4):585–95.10.1016/j.neuron.2010.04.016 [PubMed: 20510862]

42. Simon DA, Daw ND. Neural correlates of forward planning in a spatial decision task in humans. Journal of Neuroscience. 2011 forthcoming. [PubMed: 21471389]

43. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans choices and striatal prediction errors. Neuron. 2011 forthcoming. [PubMed: 21435563]

44•• . van der Meer M, Redish A. Covert expectation-of-reward in rat ventral striatum at decision points. Frontiers in Integrative Neuroscience. 2009:3. This study examines the activity of ventral striatal neurons during hippocampal preplay at decision points in a T-maze. These neurons signaled expectation of future reward at choice points, periods during which hippocampal signals have been found to preferentially reactivate candidate paths in the course of deliberation. These reward reactivations decreased with experience (and performance), consistent with accounts of reduced deliberation over experience, and suggestive of a transfer from deliberative, putatively goal-directed, processes to more automatic, habitual control.

45. van der Meer Maa Johnson A, Schmitzer-Torbert NC, Redish aD. Triple Dissociation of Information Processing in Dorsal Striatum, Ventral Striatum, and Hippocampus on a Learned Spatial Decision Task. Neuron. 2010;67(1):25–32.10.1016/j.neuron.2010.06.023

46. Johnson A, Redish AD. Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. Neural Networks. 2005;18(9):1163–71.10.1016/j.neunet.2005.08.009 [PubMed: 16198539]

47. Foster D, Wilson M. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature. 2006;440(7084):680–3. [PubMed: 16474382]

48. Diba K, Buzsáki G. Forward and reverse hippocampal place-cell sequences during ripples. Nature neuroscience. 2007;10(10):1241–2. [PubMed: 17828259]

49. Peyrache A, Khamassi M, Benchenane K, Wiener S, Battaglia F. Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. Nature Neuroscience. 2009;12(7):919–26. [PubMed: 19483687]

50. Gupta A, van der Meer M, Touretzky D, Redish A. Hippocampal replay is not a simple function of experience. Neuron. 2010;65(5):695–705. [PubMed: 20223204]

51. Kimchi EY, Laubach M. The dorsomedial striatum reflects response bias during learning. The Journal of neuroscience. 2009;29(47):14891–902.10.1523/JNEUROSCI.4060-09.2009 [PubMed: 19940185]

52. Kimchi EY, Laubach M. Dynamic encoding of action selection by the medial striatum. Journal of Neuroscience. 2009;29(10):3148–59.10.1523/JNEUROSCI.5206-08.2009 [PubMed: 19279252]

53••. Thorn, Ca; Atallah, H.; Howe, M.; Graybiel, AM. Differential Dynamics of Activity Changes in Dorsolateral and Dorsomedial Striatal Loops during Learning. Neuron. 2010;66(5):781–95. This study examined simultaneous neural activity in DMS and DLS while rodents learned a conditional T-maze task. Across the population, DMS activity reached peak structure midway through the task, while DLS structure continued to increase. The authors suggest that DLS patterns do not effect behavioral control until DMS activity becomes less prevalent, relative to DLS. 10.1016/j.neuron.2010.04.036 [PubMed: 20547134]

54. Yin HH, Costa RM. Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. Nature Neuroscience. 2009;12(3):333–42.10.1038/nn.2261 [PubMed: 19198605]

55. Frank M, Doll B, Oas-Terpstra J, Moreno F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nature Neuroscience. 2009;12(8):1062–8. [PubMed: 19620978]
56. Frank M, Hutchison K. Genetic contributions to avoidance-based decisions: striatal D2 receptor polymorphisms. Neuroscience. 2009;164(1):131–40. [PubMed: 19393722]
57. Voorn P, Vanderschuren LJMJ, Groenewegen HJ, Robbins TW, Pennartz CMa. Putting a spin on the dorsal-ventral divide of the striatum. Trends in neurosciences. 2004;27(8):468–74.10.1016/j.tins.2004.06.006 [PubMed: 15271494]
58. Paxinos, G.; Watson, C. The rat brain in stereotaxic coordinates. Academic Press; 2007.
59. Bromberg-Martin ES, Matsumoto M, Nakahara H, Hikosaka O. Multiple Timescales of Memory in Lateral Habenula and Dopamine Neurons. Neuron. 2010;67(3):499–510.10.1016/j.neuron.2010.06.031 [PubMed: 20696385]
60. Aldridge J, Tindell A, Berridge K, Zhang J, Smith K, et al. A Neural Computational Model of Incentive Salience. PloS Computational Biology. 2009:5.
61. Joel D, Weiner I. The connections of the dopaminergic system in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. Neuroscience. 2000;96(3):451–74. [PubMed: 10717427]
62. Zahm DS. The evolving theory of basal forebrain functional-anatomical 'macrosystems'. Neuroscience and Biobehavioral Reviews. 2006;30(2):148–72.10.1016/j.neubiorev.2005.06.003 [PubMed: 16125239]
63. Bouret S, Richmond BJ. Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. The Journal of Neuroscience. 2010;30(25):8591–601.10.1523/JNEUROSCI.0049-10.2010 [PubMed: 20573905]
64. Shiflett M, Balleine B. At the limbic–motor interface: disconnection of basolateral amygdala from nucleus accumbens core and shell reveals dissociable components of incentive motivation. European Journal of Neuroscience. 2010:1735–1743. [PubMed: 21044174]
65. Daw N, Courville A, Dayan P. Semi-rational models of conditioning: the case of trial order. The Probabilistic Mind: Prospects for Bayesian Cognitive Science. 2008:431–52.
66. Yin HH. The Sensorimotor Striatum Is Necessary for Serial Order Learning. Journal of Neuroscience. 2010;30(44):14719–23.10.1523/JNEUROSCI.3989-10.2010 [PubMed: 21048130]
67. Botvinick MM, Niv Y, Barto AC. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition. 2009;113(3):262–80.10.1016/j.cognition.2008.08.011 [PubMed: 18926527]
68. Beck JM, Ma WJ, Kiani R, Hanks T, Churchland AK, Roitman J, et al. Probabilistic population codes for Bayesian decision making. Neuron. 2008;60(6):1142–52.10.1016/j.neuron.2008.09.021 [PubMed: 19109917]
69••. Kiani R, Shadlen M. Representation of confidence associated with a decision by neurons in the parietal cortex. science. 2009;324:5928. The authors demonstrate that, while performing a saccadic decision task using noisy sensory input, monkey lateral intraparietal neurons encode a measure of decision confidence, reflected behaviorally by the monkey's preference for a small but certain alternate option. Across the recorded population, these trials of less-confident decisions were accompanied by a more homogeneous, medial level of activity, relative to the more peaked distributions observed when the monkey waived the certain option for one of the two gamble targets.
70. Gold J, Shadlen M. Banburismus and the Brain:: Decoding the Relationship between Sensory Stimuli, Decisions, and Reward. Neuron. 2002;36(2):299–308. [PubMed: 12383783]
71. Berkes P, Orbán G, Lengyel M, Fiser J. Spontaneous Cortical Activity Reveals Hallmarks of an Optimal Internal Model of the Environment. Science. 2011;331(6013):83–7. [PubMed: 21212356]
72. Krajbich I, Armel C, Rangel A. Visual fixations and the computation and comparison of value in simple choice. Nature Neuroscience. 2010;13(10):1292–8. [PubMed: 20835253]
73. Armel K, Beaumel A, Rangel A. Biasing simple choices by manipulating relative visual attention. Judgment and Decision Making. 2008;3(5):396–403.
74. Erev I, Ert E, Yechiam E. Loss aversion, diminishing sensitivity, and the effect of experience on repeated decisions. Journal of Behavioral Decision Making. 2008;21(5):575–97.

75. Stewart N, Chater N, Brown G. Decision by sampling. Cognitive Psychology. 2006;53(1):1–26. [PubMed: 16438947]
76. Lengyel M, Dayan P. Hippocampal contributions to control: The third way. Advances in Neural Information Processing Systems. 2008;20:889–96.
77. Constantino S, Daw N. A closer look at choice. Nature Neuroscience. 2010;13(10):1153–4. [PubMed: 20877275]
78. Lansink CS, Goltstein PM, Lankelma JV, McNaughton BL, Pennartz CMa. Hippocampus leads ventral striatum in replay of place-reward information. PLoS Biology. 2009;7(8)10.1371/journal.pbio.1000173 [PubMed: 19688032]
79••. Ding L, Gold JI. Caudate Encodes Multiple Computations for Perceptual Decisions. Journal of Neuroscience. 2010;30(47):15747–59. The authors examine the activity of primate caudate neurons during the performance of a reaction time version of the random dot motion task. Their observations corroborate theories of striatal action selection as operating on noisy evidence about state identities and resulting action values. 10.1523/JNEUROSCI.2894-10.2010 [PubMed: 21106814]
80. Dayan P, Daw ND. Decision theory, reinforcement learning, and the brain. Cognitive, affective & behavioral neuroscience. 2008;8(4):429–53.10.3758/CABN.8.4.429
81. Bogacz R, Larsen T. Integration of Reinforcement Learning and Optimal Decision-Making Theories of the Basal Ganglia. Neural Computation. 2010:1–35.
82••. Rao RPN. Decision Making Under Uncertainty: A Neural Model Based on Partially Observable Markov Decision Processes. Frontiers in Computational Neuroscience. 2010;4:1–18. This paper advances a model in which cortical representations of belief states are employed by striatum during value computation and action selection. The model is used to explain previously observations of cortical activity during a random dot motion task, and offers a new explanation of uncertainty-driven activity in dopaminergic neurons in [85]. 10.3389/fncom.2010.00146
83. Law C, Gold J. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nature Neuroscience. 2009;12(5):655–63. [PubMed: 19377473]
84. Larsen T, Leslie D, Collins E, Bogacz R. Posterior weighted reinforcement learning with state uncertainty. Neural computation. 2010;22(5):1149–79. [PubMed: 20100078]
85. Nomoto K, Schultz W, Watanabe T, Sakagami M. Temporally Extended Dopamine Responses to Perceptually Demanding Reward-Predictive Stimuli. Journal of Neuroscience. 2010;30(32):10692–9. [PubMed: 20702700]
86. Bromberg-Martin ES, Hikosaka O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron. 2009;63(1):119–26.10.1016/j.neuron.2009.06.009 [PubMed: 19607797]
87. Kemp C, Tenenbaum J. The discovery of structural form. Proceedings of the National Academy of Sciences. 2008;105(31):10687–92.
88. Gershman SJ, Blei DM, Niv Y. Context, learning, and extinction. Psychological Review. 2010;117(1):197–209.10.1037/a0017808 [PubMed: 20063968]
89. Gershman SJ, Niv Y. Learning latent structure: carving nature at its joints. Current Opinion in Neurobiology. 2010;20(2):251–6.10.1016/j.conb.2010.02.008 [PubMed: 20227271]
90. Braun D, Mehring C, Wolpert D. Structure learning in action. Behavioural Brain Research. 2010;206(2):157–65. [PubMed: 19720086]

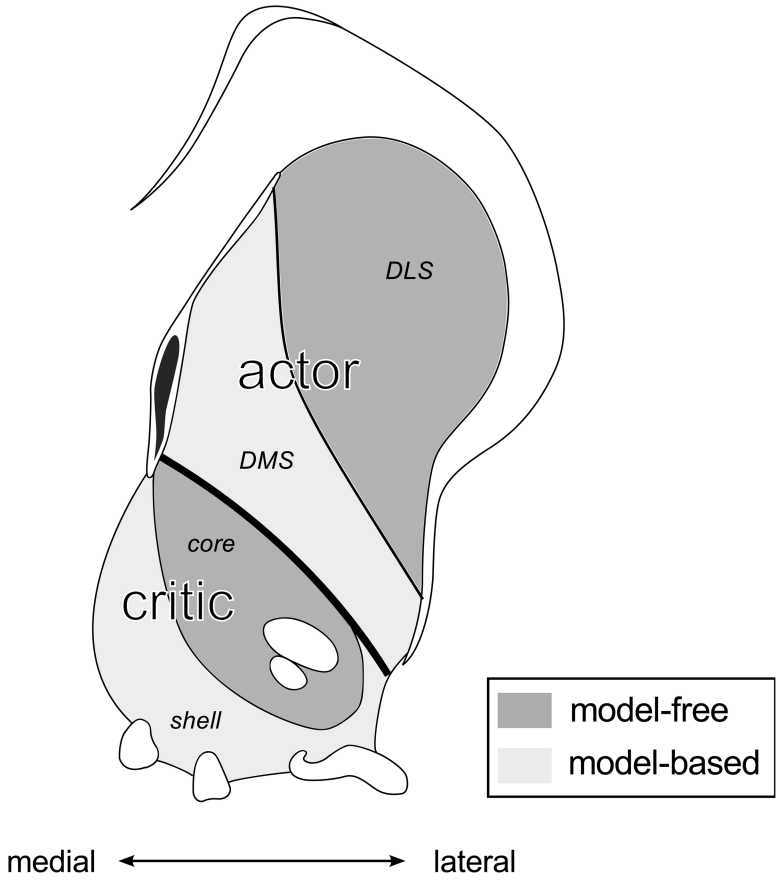

Figure 1. The dual-actor/critic framework

The dorsal/ventral divide of the actor/critic model is extended to include recent theoretical and experimental advances supporting further functional subdivisions of each region (after Yin et al. [33]). These parallel circuits implement different approaches to reinforcement learning, either "model-free" (dark grey) or "model-based" (light grey).

The dorsal region is now divided medial/lateral (though a gradient may be more accurate [57]), each supporting a different "actor" submodule: a dorsolateral area, supporting a model-free actor; and the dorsomedial region, a substrate for representations that enable model-based planning. Further, current evidence suggests that the ventral region may itself be functionally subdivided, along the boundary between nucleus accumbens "core" and "shell". These regions are each crucial for different forms of Pavlovian responding: preparatory and consummatory, respectively. Computationally, they correspond to model-based and model-free critics, computing the net present expected value of the current state using, respectively, either state-value mappings (purely based on reinforcement history) or state-outcome and outcome-value predictions derived from a world model.

Thus, model-based RL may offer a process-level description of striatal subregion function that encompasses both goal-directed instrumental and consummatory Pavlovian behaviors, paralleling the unification of habitual action and preparatory Pavolovian responses embodied in the original, model-free, actor/critic formulation.

*Schematic coronal slice of rat striatum modified with permission from Paxinos and Watson [*58].