Published in final edited form as:
Curr Opin Neurobiol. 2012 December ; 22(6): 1075–1081. doi:10.1016/j.conb.2012.08.003.

The ubiquity of model-based reinforcement learning

Bradley B Doll[a,b], **Dylan A Simon**[c], and **Nathaniel D Daw**[b,c]
[a]Department of Psychology, Columbia University, New York, NY
[b]Center for Neural Science, New York University, New York, NY
[c]Department of Psychology, New York University, New York, NY

Abstract

The reward prediction error theory of dopamine function has enjoyed great success in the neuroscience of learning and decision-making. This theory is derived from model-free reinforcement learning in which choices are made simply on the basis of previously realized rewards. Recently, attention has turned to correlates of more flexible, albeit computationally complex, model-based methods in the brain. These methods are distinguished from model-free learning by their evaluation of candidate actions using expected future outcomes according to a world model. Puzzlingly, signatures from these computations seem to be pervasive in the very same regions previously thought to support model-free learning. Here, we review recent behavioral and neural evidence about these two systems, in attempt to reconcile their enigmatic cohabitation in the brain.

Introduction

The reward prediction error (RPE) theory of dopamine (DA) [1, 2] has been a remarkably influential account of neural mechanisms for learning from reward and punishment. This marriage of computer science and neuroscience posits that dopaminergic bursts and dips transmit plasticity-modulating teaching signals to cortico-striatal circuits, training the selection of rewarded actions and avoidance of punished ones.

The computational end of this theory is known as 'model-free' reinforcement learning (RL), and derives its name from the fact that the learner need not attempt to understand the sequential transition structure of the task (a 'world model') to maximize reward. Instead of utilizing this structure to assess future outcomes, these models learn which actions are beneficial through direct experience with their rewarding consequences. This learning scheme has clear benefits in terms of computational efficiency, but there are costs as well. The capacity to learn about the world's structure, lacking in these approaches, is most beneficial (and empirically most apparent) when flexible changes in behavior are required, e.g., when goals or aspects of the environment change, necessitating rapid reevaluation.

Psychologically, model-free algorithms reinvent Thorndike's [3] early law of effect, the reinforcement principle according to which rewarded actions tend to be repeated. In many ways, the refutation of this demonstrably incomplete behaviorist principle sowed the seeds

Correspondence to: Nathaniel D Daw.

of the cognitive revolution [4]. Even rats can do more than repeat successful actions: they can, for instance, learn the layout of a maze (a prototypical example of a world model) and plan novel routes in it.

However, it is easier to demonstrate that the brain is smart than it is to understand how it manages this sophistication. Fortunately, just as model-free RL provides a theory specific enough to be refuted, the engineering literature offers a second class of algorithms that have recently been identified as promising candidates for formalizing a more flexible alternative [5, 6, 7]. Such 'model-based' RL approaches learn the sequential contingencies of events and actions in a task (which outcomes follow which actions, e.g., where different paths in a maze lead), which can be used adaptively and dynamically to compute ideal actions by simulating their consequences. It is this sequentially structured world model, and its use in forward-looking computations, that distinguish model-based from model-free RL.

Here, we review recent efforts that leverage model-based methods to uncover the neurobiological underpinnings of more flexible decision making, and to dissociate them from their putative model-free counterparts.

From goals and habits...

The model-based vs. model-free dichotomy was proposed to capture a long-standing distinction in psychology between two classes of instrumental behavior known as goal-directed and habitual [5, 8]. This distinction is operationalized with tasks that use revaluation probes, such as training an animal to lever-press for food when hungry, then testing performance when full. Revaluation interrogates whether choice of an action (lever-pressing) is affected by consideration of its outcome (the food, now worthless) or merely determined by previous reinforcement (which occurred when hungry). If behavior instantly adjusts to reflect the new value of the outcome that the action would obtain, this demonstrates that the choice is 'goal-directed', i.e., derived from a representation of the action's specific consequences, or in RL terms, model-based. Insensitivity to revaluation instead indicates 'habits', choices made without regard for any representation of outcome identity but instead selected based on previously realized value, as with model-free RL.

In experiments, each account prevails under different circumstances: animals can either respect or ignore the new outcome value. Together with lesions dissociating these functions (below) and the strong, pre-existing theory associating DA and model-free RL, these revaluation results suggested that the brain contained separate, competing systems for model-based and model-free RL [5]. This view also offers a theoretical explanation for when each approach should dominate. Given estimates from both systems, one favorable strategy is to select the least uncertain among them [5]. A recent theory framed arbitration more explicitly in terms of the costs (time) and benefits (better reward harvesting) of performing model-based evaluation [9]. The newer formulation has stronger decision-theoretic foundations and better process-level plausibility, but the accounts are similar in effect because the reward increment expected for model-based evaluation depends on uncertainty. Thus both frameworks correctly predict many circumstances in which putatively model-free behavior dominates, including after overtraining [10] and in conditions when action-reward contiguity is low [11] or reward variability is high [12], as well as one direct test of uncertainty's role in arbitration [13]. The cost/benefit analysis (specifically, the time cost of model-based evaluation) may also explain why chronic stress in rats [14] and acute stress in humans [15, 16] shift both species toward habitual behavior.

The revaluation assay can dissociate not only behavior but also its neural underpinnings in rats [8]. Recently, this approach has been extended to humans, suggesting conservation of brain and behavior across species [17]. Lesions in the rodent indicate that integrity of the

dorsolateral striatum is key for the formation of habitual behavior [18]. The role of (at least this portion of) striatum resonates with the model-free RPE theory of its phasic DA input, which knockout and optogenetic studies demonstrate is necessary for some measures of behavioral conditioning [19, 20].

The more puzzling counterpart to these results, however, is that goal-directed (putatively model-based) behavior requires adjacent dorsomedial striatum [21]. It is difficult to reconcile the structural similarity of these parts of striatum with the substantial differences in their theorized computations. Thus, these results paint a confusing picture, whereby a dopamine-rich cortico-striatal loop, analogous to that commonly thought to implement model-free RL, appears to be involved additionally in model-based RL. More particularly, standard model-based RL has no use for an RPE of the sort usually associated with striatum's dopamine input, and instead uses quite different teaching signals [22]. One route to a solution might be potential differences in the properties of the areas' dopamine inputs. Indeed, different striatal regions are irrigated by dissociable groups of midbrain DA neurons [23]. However, relatively few electrophysiological recordings have suggested systematic variation in signaling properties across them [24, 25], and the interpretation of such variation can be further complicated by challenges in identifying which cells are dopaminergic [26].

A similarly mixed picture arises from studies investigating dopamine's causal effects in revaluation tasks. Indeed, habits can be induced using DA-agonizing drugs of abuse in place of natural rewards like food [27]; while for natural rewards, deafferention of DA cells targeting dorsolateral striatum [28], and deletion of NMDA receptors in DA neurons in knockout mice [29] prevent transition from goal-directed to habitual behavior. Conversely, pharmacological blockade of DA has no effect on measures of goal-directed instrumental learning [30] at doses that affect (presumably model-free) Pavlovian conditioning. On the other hand, recent results using a human task inspired by traditional associative learning models of the revaluation paradigm seem to suggest a counterintuitive role for DA in goal-directed rather than habitual behavior [31, 32]. One possibility is that these results reflect dopamine's involvement in (and goal-directed behavior's reliance on) prefrontal cognitive functions besides reinforcement, such as working memory. However, unlike the standard devaluation paradigm, the central manipulation of this task — involving different combinations of the same stimuli as both cues and outcomes for responses — is not clearly interpretable in terms of model-based and model-free computations, so these results may not easily generalize.

...to model-based and model-free RL

Indeed, model-based vs. model-free, as the key dimension distinguishing learning strategies, extends beyond its proposed operationalization in goal-directed and habitual behaviors. The computational distinction applies also, for instance, in Pavlovian conditioning, spatial navigation, and other cognitive problems, especially those from the engineering literature where it arose. Accordingly, numerous recent studies, mostly using fMRI, have aimed explicitly to dissociate model-based from model-free reinforcement learning using learning tasks inspired by the computational RL literature. These can roughly be grouped into two classes. One is sequential decision tasks — mazes or more abstract multistep sequences — in which model-based methods can learn the sequential transition structure and leverage it to evaluate actions (Figure 1)[22, 33, 34, 35, 36, 12, 37, 38, 39].The second involve explicit or implicit counterfactual structure, where information about rewards not actually received can be inferred or observed [40, 41, 42, 13, 43, 44]. A typical example is a serial reversal contingency, where a drop in the value of one option implies an increase in the other's value. Purely reinforcement-based model-free RL would be blind to such structure. Note, however, that while such tasks go beyond model-free RL, they don't as directly exercise the

key affirmative features of model-based RL as we have defined it, i.e. the computation of values using a sequential transition model of an action's consequences.

From both sorts of studies, the overall sense is that model-based influences appear ubiquitous more or less wherever the brain processes reward information. The most expected of these influences are widespread reports about model-based value signals in ventromedial prefrontal cortex (vmPFC) and adjacent orbitofrontal cortex (OFC), which have previously been identified with goal-directed behavior using devaluation tasks [45, 46]. vmPFC has been proposed to be the human homologue of rat prelimbic cortex, which is required for goal-directed behavior [8]. OFC is also implicated in model-based Pavlovian valuation in rats and goal values in monkeys [47, 48], though understanding this area across species and methods is plagued by multiple factors [49]. More unexpectedly, several reports now indicate that RPE correlates in the ventral striatum — long thought to be a human counterpart to the DA response and thus a core component of the putative model-free system — also show model-based influences [33, 34, 44]. Even DA neurons, the same cells that launched the model-free theories due to their RPE properties [1, 2], communicate information not available to a standard model-free learner [41].

The harder part of this hunt, then, seems to be for neural correlates of exclusively model-free signals, which are surprisingly sparse given the prominence of the model-free DA accounts. The most promising candidate may be a region of posterior putamen that has been implicated in extensively trained behavior in a habit study [17] and a sequential decision task [37], and may correspond to the dorsolateral striatal area associated with habits in rodents [18]. The foundation of both fMRI results, however, was overtraining (a classic promoter of habits), rather than whether these areas reflect values learned or updated by model-free methods. Indeed, value correlates in a nearby region of putamen have been reported to follow model-based rather than model-free updating using the computational definition [34].

A different, promising path for isolating model-based RL is neural correlates related to the model itself. Representations of anticipated future states or out-comes — rather than just their consequences for reward — are what defines model-based RL. Hippocampal recordings in the rat have shown evidence of forward model 'lookahead sweeps' to candidate future locations at maze choice points [35]. These data fit well with the spatial map-encoding properties of hippocampus [50], and may permit striatum to signal value for simulated rather than actually experienced outcomes [36]. Hippocampus is similarly implicated in a study that examines learning predictive world models outside the context of a decision task [51]. A number of cortical areas have also been observed with correlates related to model learning [22] and evaluation [34]. Work in learning tasks with task-relevant hidden structure may also speak to the construction of world models [52]. These studies have implicated the lateral PFC, an area often associated with working memory function, in discovery of this structure.

The way forward

Why the ubiquity of evidence for model-based RL? Several factors probably contribute, all of which point to important opportunities for progress.

First, the resolution and other limitations of the BOLD signal may conceal distinctions that would be visible using more invasive techniques. The explosion of human studies explicitly examining model-based RL is recent enough that analogous animal electrophysiological studies are largely not yet available. A strength of fMRI is to develop tasks and analyses, and to locate areas for further study; the time is now ripe to test similar tasks in animals. For instance, further dissociations might be found in the signaling properties of different DA cell

groups and projections [23], potentially elucidating the potential co-expression of model-based and model-free signals by DA cells.

Second, the brain's RPE systems may be smarter than they have been made out to be, yet still essentially model-free. Of the different characteristics that have been taken as hallmarks of model-based RL, some are easier than others to accommodate in a lightly modified model-free system. In particular, a model-free learner can generalize learning from one state to another, without additional experience, if its inputs corresponding to those states overlap (Figure 2). This is a particularly plausible explanation for seemingly model-based inference in serial reversal and similar tasks [40, 41, 44], as indeed the authors of some of these studies have pointed out. If counterfactual updating in these tasks occurs implicitly, due to generalization, then it wouldn't involve forward modeling of future states. In this respect, tasks involving sequential contingencies are stronger and more canonical tests of model-based RL, but variants of the same representational trick can in principle apply even there [53]. For instance, if actions are represented in terms of their associated outcomes – for instance if the representation for a lever that produces food overlaps with that for the food itself – and if these inputs (themselves now, in effect, a sort of world model) are mapped to values using even model-free RL, then the learned value will be substantially shared between the lever and the food. In this case, if the food is devalued, the leverpress value will also decline immediately, and the resulting behavior will appear goal directed. This approach might help to explain the involvement of similar striatal circuits in both goal-directed and habitual behavior. More elaborate versions of this scheme can apply to arbitrary sequential tasks, but such a strategy is easier to spot, and potentially to rule out, in tasks with deeper sequential structure and changing transition contingencies [34].

Third, there may be hitherto unanticipated crosstalk or integration between model-based and model-free systems. It is reasonable to imagine that model-based capacities evolved on top of an earlier model-free system, rather than separately and in parallel. One related proposal that has received some recent support [54, 55] is that explicit representations (putatively PFC and hippocampus dependent) can directly bias an underlying model-free learner, permitting higher order contingency information to be learned by the model-free architecture [56]. A model-based system in the brain might similarly leverage a model-free learner, as with some model-based algorithms that incorporate model-free quantities in order to reduce computational overhead [57, 58, 59]. Different modes of behavior may simply reflect different aspects of a more complex, integrated learning system. For example, there is evidence for one type of model-based learning that embeds a Pavlovian system, giving rise to some aspects of both sorts of behavior simultaneously [60].

The explosion of model-free RL approaches in psychology and neuroscience has led to tremendous progress over the past 15 years. Model-based approaches hold a similar promise, but with their complexity comes the need for close attention to the specific computations and predictions these models make, and a re-evaluation of their relationship with established model-free approaches.

Acknowledgments

The authors are supported by Scholar Awards from the McKnight foundation and the McDonnell foundation, NIMH grant 1R01MH087882-01 and NINDS grant 1R01NS078784-01. We thank Sara Constantino for figure illustration.

References

1. Barto, AG. Adaptive critics and the basal ganglia. In: Houk, JC.; Davis, JL.; Beiser, DG., editors. Models of information processing in the basal ganglia. Cambridge, MA: MIT Press; 1995. p. 215-232.Ch. xii
2. Montague PR, Dayan P, Sejnowski TJ. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci. 1996; 16(5):1936–1947. [PubMed: 8774460]
3. Thorndike EL. Animal intelligence: An experimental study of the associative processes in animals. Psychological Review Monograph Supplement. 1898; 2(4):1–8.
4. Tolman EC. Cognitive maps in rats and men. Psychol Rev. 1948; 55:189–208. [PubMed: 18870876]
5. Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8(12):1704–1711. [PubMed: 16286932]
6. Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. 1999; 12(7–8):961–974. [PubMed: 12662639]
7. Solway A, Botvinick MM. Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates. Psychol Rev. 2012; 119(1):120–154. [PubMed: 22229491] • This paper reframes model-based value computation in terms of Bayesian probabilistic inference, linking model-based RL with contemporary Bayesian inference theories in psychology and neuroscience.
8. Balleine BW, O'Doherty JP. Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 2009; 35(1):48–69. [PubMed: 19776734]
9. Keramati M, Dezfouli A, Piray P. Speed/accuracy trade-off between the habitual and the goal-directed processes. PLoS Comput Biol. 2011; 7(5):e1002055. [PubMed: 21637741] • This paper frames the distinction between goal-directed and habitual behavior as a speed/accuracy tradeoff in reward attainment. In particular, the cost of utilizing the slow but more accurate model-based system to decide is compared to the benefit in reward accumulation this computation would achieve. If the benefit exceeds the cost, the model-based system is used, otherwise the fast but less accurate model-free system continues to control actions. In reward revaluation tasks, the benefit of the model-based computation decreases as the different values of the possible actions are learned, allowing the model-free system to dominate and habitual behavior to be observed.
10. Adams CD. Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q J Exp Psychol B. 1982; 34(2):77–98.
11. Derusso AL, Fan D, Gupta J, Shelest O, Costa RM, Yin HH. Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. Front Integr Neurosci. 4
12. Simon DA, Daw ND. Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K. Environmental statistics and the trade-off between model-based and TD learning in humans. Advances in Neural Information Processing Systems. 2011; 24:127–135. • A behavioral study using a sequential decision task shows that the relative contribution of model-based and model-free learning can be manipulated by adjusting the statistics of rewards. These results support the predictions of computational theories of the tradeoff between the controllers.
13. Beierholm UR, Anen C, Quartz S, Bossaerts P. Separate encoding of model-based and model-free valuations in the human brain. Neuroimage. 2011; 58(3):955–962. [PubMed: 21757014]
14. Dias-Ferreira E, Sousa JaC, Melo I, Morgado P, Mesquita AR, Cerqueira JaJ, Costa RM, Sousa N. Chronic stress causes frontostriatal reorganization and affects decision-making. Science. 2009; 325(5940):621–625. [PubMed: 19644122]
15. Schwabe L, Wolf OT. Socially evaluated cold pressor stress after instrumental learning favors habits over goal-directed action. Psychoneuroendocrinology. 2010; 35(7):977–986. [PubMed: 20071096]
16. Schwabe L, Wolf OT. Stress-induced modulation of instrumental behavior: From goal-directed to habitual control of action. Behav Brain Res. 2011; 219(2):321–328. [PubMed: 21219935]
17. Tricomi E, Balleine BW, O'Doherty JP. A specific role for posterior dorsolateral striatum in human habit learning. Eur J Neurosci. 2009; 29(11):2225–2232. [PubMed: 19490086]

18. Yin HH, Knowlton BJ, Balleine BW. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci. 2004; 19(1):181–189. [PubMed: 14750976]

19. Tsai H-C, Zhang F, Adamantidis A, Stuber GD, Bonci A, de Lecea L, Deisseroth K. Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. Science. 2009; 324(5930):1080–1084. [PubMed: 19389999]

20. Zweifel LS, Parker JG, Lobb CJ, Rainwater A, Wall VZ, Fadok JP, Darvas M, Kim MJ, Mizumori SJY, Paladini CA, Phillips PEM, Palmiter RD. Disruption of NMDAR-dependent burst firing by dopamine neurons provides selective assessment of phasic dopamine-dependent behavior. Proc Natl Acad Sci. 2009; 106(18):7281–7288. [PubMed: 19342487]

21. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci. 2005; 22(2):513–523. [PubMed: 16045504]

22. Glascher J, Daw N, 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): 585–595. [PubMed: 20510862]

23. Haber SN, Knutson B. The reward circuit: Linking primate anatomy and human imaging. Neuropsychopharmacology. 2009; 35(1):4–26. [PubMed: 19812543]

24. Matsumoto M, Hikosaka O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature. 2009; 459(7248):837–841. [PubMed: 19448610]

25. Brischoux F, Chakraborty S, Brierley DI, Ungless MA. Phasic excitation of dopamine neurons in ventral vta by noxious stimuli. Proc Natl Acad Sci U S A. 2009; 106(12):4894–4899. [PubMed: 19261850]

26. Ungless MA, Grace AA. Are you or aren't you? challenges associated with physiologically identifying dopamine neurons. Trends Neurosci. 2012; 35(7):422–430. [PubMed: 22459161]

27. Zapata A, Minney VL, Shippenberg TS. Shift from goal-directed to habitual cocaine seeking after prolonged experience in rats. J Neurosci. 2010; 30(46):15457–15463. [PubMed: 21084602]

28. Faure A, Haberland U, Coné F, Massioui NE. Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. J Neurosci. 2005; 25(11):2771–2780. [PubMed: 15772337]

29. Wang LP, Li F, Wang D, Xie K, Wang D, Shen X, Tsien JZ. Nmda receptors in dopaminergic neurons are crucial for habit learning. Neuron. 2011; 72(6):1055–1066. [PubMed: 22196339]

30. Dickinson A, Smith J, Mirenowicz J. Dissociation of pavlovian and instrumental incentive learning under dopamine antagonists. Behav Neurosci. 2000; 114(3):468–483. [PubMed: 10883798]

31. de Wit S, Barker RA, Dickinson AD, Cools R. Habitual versus goal-directed action control in Parkinson disease. J Cogn Neurosci. 2011; 23(5):1218–1229. [PubMed: 20429859]

32. de Wit S, Standing HR, Devito EE, Robinson OJ, Ridderinkhof KR, Robbins TW, Sahakian BJ. Reliance on habits at the expense of goal-directed control following dopamine precursor depletion. Psychopharmacology (Berl). 2012; 219(2):621–631. [PubMed: 22134475]

33. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on humans' choices and striatal prediction errors. Neuron. 2011; 69(6):1204–1215. [PubMed: 21435563] • fMRI study utilizing a sequential decision task and a novel analysis to dissociate model-based from model-free learning strategies. Rather than only identifying model-based RL, the analysis can identify different affirmative signatures of both model-based and model-free RL. Evidence for both strategies' contributions is reported in behavior and in RPE signals in ventral striatum.

34. Simon DA, Daw ND. Neural correlates of forward planning in a spatial decision task in humans. J Neurosci. 2011; 31(14):5526–5539. [PubMed: 21471389]

35. 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. [PubMed: 20624589] • This paper compares recordings from hippocampus, ventral, and dorsal striatum in different rats solving a multiple-T maze. The authors found evidence for hippocampal "lookahead" sweeps, ventral striatal reward activation at reward site at a choice point prior to reward, and dorsal striatal activation at multiple choice points. These findings suggest the co-existence of model-based and model-free computations in the rat brain.

36. van der Meer MAA, Redish AD. Expectancies in decision making, reinforcement learning, and ventral striatum. Front Neurosci. 2010; 4:6. [PubMed: 21221409]
37. Wunderlich K, Dayan P, Dolan RJ. Mapping value based planning and extensively trained choice in the human brain. Nat Neurosci. 2012; 15(5):786–791. [PubMed: 22406551] • This study dissociated striatal loci of forward modeled and extensively trained action values. These results are suggestive of homology between human brain studied here and rat brain as assessed in classic devaluation tasks.
38. Fermin A, Yoshida T, Ito M, Yoshimoto J, Doya K. Evidence for model-based action planning in a sequential finger movement task. J Mot Behav. 2010; 42(6):371–379. [PubMed: 21184355] • A novel and promising behavioral task is used to demonstrate evidence for model-based planning in humans. The task requires subjects to learn different mappings from buttons to directions of movement; model-based RL can exploit this structure to plan arbitrary trajctories. Model-based planning is demonstrated by the way accuracy is affected by experience with the button mapping and by planning time.
39. Pan X, Sawa K, Tsuda I, Tsukada M, Sakagami M. Reward prediction based on stimulus categorization in primate lateral prefrontal cortex. Nat Neurosci. 2008; 11(6):703–712. [PubMed: 18500338]
40. Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci. 2006; 26(32):8360–8367. [PubMed: 16899731]
41. Bromberg-Martin ES, Matsumoto M, Hong S, Hikosaka O. A pallidushabenula-dopamine pathway signals inferred stimulus values. J Neurophysiol. 2010; 104(2):1068–1076. [PubMed: 20538770]
42. Abe H, Lee D. Distributed coding of actual and hypothetical outcomes in the orbital and dorsolateral prefrontal cortex. Neuron. 2011; 70(4):731–741. [PubMed: 21609828]
43. Wunderlich K, Symmonds M, Bossaerts P, Dolan RJ. Hedging your bets by learning reward correlations in the human brain. Neuron. 2011; 71(6):1141–1152. [PubMed: 21943609]
44. Wimmer GE, Daw ND, Shohamy D. Generalization of value in reinforcement learning by humans. Eur J Neurosci. 2012; 35(7):1092–1104. [PubMed: 22487039]
45. Valentin VV, Dickinson A, O'Doherty JP. Determining the neural substrates of goal-directed learning in the human brain. J Neurosci. 2007; 27(15):4019–4026. [PubMed: 17428979]
46. de Wit S, Corlett PR, Aitken MR, Dickinson A, Fletcher PC. Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. J Neurosci. 2009; 29(36):11330–11338. [PubMed: 19741139]
47. McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci. 2011; 31(7):2700–2705. [PubMed: 21325538]
48. Padoa-Schioppa C, Assad JA. Neurons in the orbitofrontal cortex encode economic value. Nature. 2006; 441(7090):223–226. [PubMed: 16633341]
49. Wallis JD. Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat Neurosci. 2012; 15(1):13–19. [PubMed: 22101646]
50. O'Keefe, J.; Nadel, L. The hippocampus as a Cognitive Map. Oxford: Clarendon Press; 1978.
51. Bornstein AM, Daw ND. Dissociating hippocampal and striatal contributions to sequential prediction learning. Eur J Neurosci. 2012; 35(7):1011–1023. [PubMed: 22487032]
52. Frank MJ, Badre D. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. Cereb Cortex. 2012; 22(3):509–526. [PubMed: 21693490]
53. Dayan P. Improving generalization for temporal difference learning: The successor representation. Neural Comput. 1993; 5(4):613–624.
54. Doll BB, Hutchison KE, Frank MJ. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. J Neurosci. 2011; 31(16):6188–6198. [PubMed: 21508242]
55. Biele G, Rieskamp J, Krugel LK, Heekeren HR. The neural basis of following advice. PLoS Biol. 2011; 9(6):e1001089. [PubMed: 21713027]
56. Doll BB, Jacobs WJ, Sanfey AG, Frank MJ. Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res. 2009; 1299:74–94. [PubMed: 19595993]

57. Simon, DA.; Daw, ND. Dual-system learning models and drugs of abuse. In: Gutkin, B.; Ahmed, SH., editors. Computational Neuroscience of Drug Addiction, Vol. 10 of Computational Neuroscience. Springer; 2012. p. 145-162.Ch. 5

58. Sutton, RS. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. Proceedings of the Seventh International Conference on Machine Learning, GTE Laboratories Incorporated, Morgan Kaufmann; 1990. p. 216-224.

59. Walsh, TJ.; Goschin, S.; Littman, ML. Integrating sample-based planning and model-based reinforcement learning. AAAI Conference on Artificial Intelligence; 2010.

60. Huys QJM, Eshel N, O'Nions E, Sheridan L, Dayan P, Roiser JP. Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. PLoS Comput Biol. 2012; 8(3):e1002410. [PubMed: 22412360] • Demonstrates how a model-based system interfaces with a putatively model-free Pavlovian system to prune decision trees and simplify forward modeling.

Highlights

- Model-free RL is a successful theory of corticostriatal DA function.
- Flexible model-based RL methods offer to enrich understanding of brain and behavior.
- Data suggest extensive overlap between putative neural correlates of these RL systems.

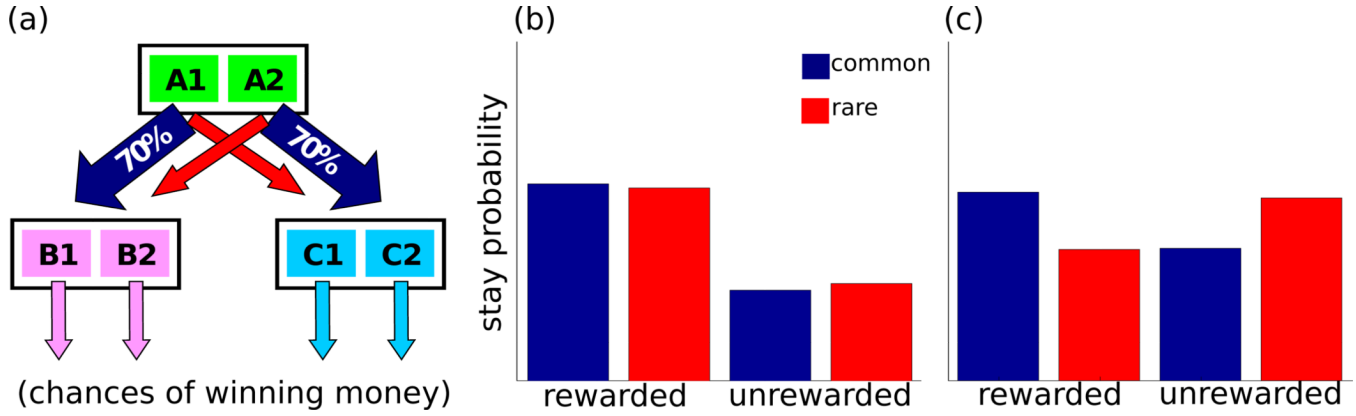

Figure 1.

Sequential task dissociating model-based from model-free learning. (**a**) A two-step decision making task [33], in which each of two two options (A1, A2) at a start state leads preferentially to one of two subsequent states (A1 to B, A2 to C), where choices (B1 vs. B2 or C1 vs C2) are rewarded stochastically with money. (**b**, **c**) Model-free and model-based RL can be distinguished by the pattern of staying vs switching of a top level choice following bottom level winnings. A model-free learner like TD(1) (**b**), tends to repeat a rewarded action without regard to whether the reward occurred after a common transition (blue, like A1 to B) or a rare one (red). A model-based learner (c) evaluates top-level actions using a model of their likely consequences, so that reward following a rare transition (e.g., A1 to C) actually increases the value of the unchosen option (A2) and thus predicts switching. Human subjects in [33] exhibited a mixture of both effects.

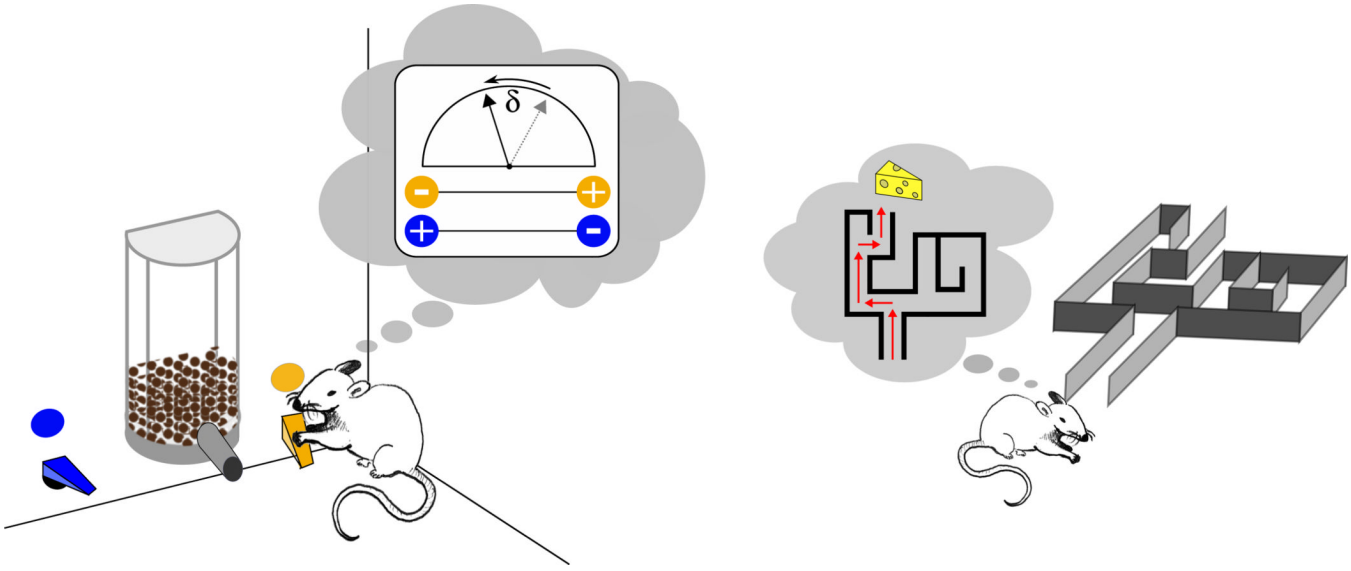

Figure 2.

Learning through value generalization (left) and model-based forward planning (right). In a reversal learning task (left), the rat has just taken an action (lever-press) and received no reward, and so updates its internal choice value representation to decrement the chosen value's option. Because the unchosen value's option is represented on the same scale, inverted, it is implicitly incremented as well. Implemented this way, learning relies on model-free updating over a modified input, and does not involve explicitly constructing or evaluating a forward model of action consequences. In a model-based RL approach to a maze task (right), the rat has an internal representation of the sequential structure of the maze, and uses it to evaluate a candidate route to the reward.