Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Reinforcement learning with Marr

Yael Niv and Angela Langdon

To many, the poster child for David Marr's famous three levels of scientific inquiry is reinforcement learning — a computational theory of reward optimization, which readily prescribes algorithmic solutions that evidence striking resemblance to signals found in the brain, suggesting a straightforward neural implementation. Here we review questions that remain open at each level of analysis, concluding that the path forward to their resolution calls for inspiration across levels, rather than a focus on mutual constraints.

Address
Psychology Department & Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, United States

Corresponding author: Niv, Yael (yael@princeton.edu)

Current Opinion in Behavioral Sciences 2016, **11**:67–73

This review comes from a themed issue on **Computational modelling**

Edited by **Peter Dayan** and **Daniel Durstewitz**

http://dx.doi.org/10.1016/j.cobeha.2016.04.005

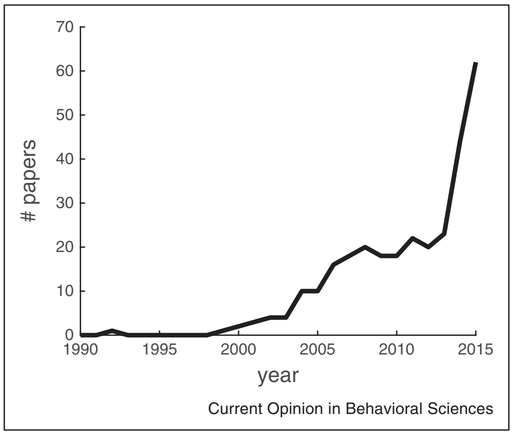
Over the past 25 years, reinforcement learning (RL) has risen from relative obscurity to scientific stardom (Figure 1), now encompassing hundreds of researchers in disciplines as varied as economics, computer science, robotics, psychology, ethology, and neuroscience. Arguably this success can be attributed to the fact that, as a field, RL straddles all three levels of Marr's famous framework of scientific inquiry in computational neuroscience [1]. At the *computational* level, RL defines a small set of normative targets (accurately predicting the sum of future rewards, choosing actions that maximize reward attained, etc.). The *algorithmic* level — a host of solutions that achieve these normative goals — elegantly derives directly from the definition of the computational targets [2]. In particular, describing decision making problems in terms of Markov (memoryless) decision processes allows for recursive computation of both reward predictions and action values, using local prediction errors [3,4]. Finally, at the *implementational* level, these algorithms have been closely tied to neural substrates of learning and prediction in the basal ganglia [5–8], and in particular, prediction errors have been linked to dopaminergic signaling in the brain [9–11]. At the risk of drawing boundaries that are sometimes artificial, here we summarize at each of these levels recent findings and current open questions in the field of RL as it is studied in the fields of psychology and neuroscience. Marr's levels, as an organizing principle, serve to highlight conceptual differences between questions asked at each level, and how findings at one level can inspire progress in another.

The computational level: the goals of a decision-making system

At the computational level, the basic goal of an agent or a decision-making system is to maximize reward and minimize punishment. Although one might argue whether this is the true goal of agents from an evolutionary perspective, different definitions of reward and punishment allow considerable flexibility. Indeed, work in recent years has elaborated on what constitutes a reward — in addition to the obvious food and shelter (and their associated conditioned reinforcers) there seem to be other forms of reward that are perhaps similarly primary in nature [12]. For instance, 'curiosity' can be seen as motivated by the goal of seeking information [13,14], and work on intrinsic motivation [15] has suggested that agents may maximize not only the sum of future rewards, but also the reduction of uncertainty about rewards in the environment [16–18]. Moreover, behavioral results, and corresponding neural recordings in monkeys, have convincingly shown that advance information is valuable in of itself, that is, even if this information cannot be acted upon [19–23].

A related line of work has asked what fictitious, internal rewards an animal (or experimenter) could design, that would assist in ultimately achieving highest fitness in the environment. In this framework of 'shaping rewards' the computational-level question is: what is the optimal (pseudo-)reward function with which learning with a specific (possibly limited) set of algorithms would end up maximizing real reward or evolutionary fitness? Recent findings have highlighted that separating the evaluative role of reward functions from their policy-shaping role is beneficial for agents that are bounded (e.g., in terms of the accuracy of their representation of the environment, their capacity for planning, or the learning algorithms they are restricted to use). That is, for different statistics of environments and structure of agents, there exist internal reward functions that lead to faster learning and higher asymptotic behavior, and these are different from the objective reward function [24,25]. By reinforcing behaviors such as exploration or information seeking that are only indirectly related to objective fitness these internal reward functions mitigate the boundedness of the agent [26], although one might argue that this is due to the reward function providing the agent with information

Figure 1



Number of papers containing the term 'reinforcement learning' published by Nature Publishing Group between 1990 and 2015.

(e.g., from past history) that was otherwise unavailable to it. In general, these questions about what constitutes a reward link back to work in the 1940s by Hull [27], with the advent of algorithms and neuroscientific tools that allow precise quantification of the reward value of different events resurrecting hitherto unanswerable questions [28•,29]. Moreover, this question now has practical import, for instance in mobile-health applications and other adaptive interventions where an optimal pseudo-reward function can promote adherence to health recommendations (e.g., exercise regimens) toward the long-term goal of improved health [30,31].

One relatively neglected computational-level goal for animals is to define (and, algorithmically, learn) an optimal representation for a task. Different representations of the same task can render it easy (in the case of a small number of Markov states, or with states that allow a smooth value function), hard (when unnecessary detail is represented in the states), verging on intractable (for non-Markov partially-observable representations) or even impossible to solve (if information that is critical to task performance is not included in the representation). Different state representations also give rise to values or policies that generalize differently to new tasks [32]. Therefore, alongside the normative goal of maximizing reward, one might define a (subsidiary) goal of optimizing task representation. Although RL applications often use a predesigned representation, in the animal (and human) kingdom most task representations are not 'given' but rather must be learned through experience, raising the algorithmic question of how this computational goal can be achieved [33]. The advent of principled, statistical methods for studying trial-by-trial learning dynamics [34] has allowed critical examination of the early phases of learning, when animals learn the 'rules of the game' (or the state representation). Behavioral findings point to a process akin to Bayesian inference in which animals attempt to use observed information to infer the unobservable (latent) causal structure of the task, which is then used to craft task states that accurately describe the task dynamics [35–38]. This 'representation learning' process has been linked to memory processes [39,40] as well as neural selective attention mechanisms [41]. As is the case here, computational level questions in RL link intimately with the algorithmic level, to which we now turn.

The algorithmic level: multiple solutions to the decision-making problem

Given the computational goal of maximizing reward, how does a decision-making agent learn which states of the world predict reward, and what actions enable their attainment? RL provides multiple algorithmic solutions to the problem of credit assignment (i.e., correctly assigning credit or laying blame for an outcome on preceding actions or states). Many of these algorithms proceed through the incremental update of state- and action-specific 'values' defined as the (discounted) sum of future expected rewards. This update hinges on the calculation of reward prediction errors that compare the predicted value to the current outcome plus expected future value [4].

A major focus in recent years has been uncovering the neural substrates of computations of the economic value of goods [42]. Here, a major algorithmic question is whether reward is indeed evaluated in terms of common currency that allows maximization over bundles of outcomes. This idea, central to economics and especially to the field of neuroeconomics, has found some support [43–45]. However, it is not yet clear that common-currency values are at all necessary in order to achieve optimality. Alternative methods for RL that hail from engineering and computer science search instead in the space of behavioral policies and can find the best policy without computing the reward value of different actions [46–48]. Indeed, recent evidence suggests that prediction error signals in the brain correspond to value differences between actions, which can supply the appropriate gradient for such 'policy search' algorithms [49].

In addition to learning values, a ubiquitous aspect of real-world tasks is time: learning *when* a reward will occur is sometimes as important as learning that it will occur. Indeed, at the heart of RL algorithms is the temporal-difference method that compares predictions across consecutive time points in a task. However, one might argue that time has been mistreated in RL, typically simplified and discretized in order to conform to the Markov assumption [50]. For example, in 'tapped delay line' or 'serial compound' representations, the passage of time is modeled using discrete sequential states that each accrue a separate value. This redundancy of values requires additional mechanisms to accelerate learning (e.g., eligibility traces [2]), and results in erroneous predictions when outcomes arrive earlier than expected [51]. Moreover, discretizing time raises the question of an appropriate resolution, and is hard-pressed to account for well-established phenomena such as scalar timing noise. Recent work attempting to address these issues proposed distributed temporal representations that span multiple durations, thereby allowing multi-timescale learning in an elegant way [52,53]. Another promising alternative is to model learning using a semi-Markov framework where (continuous) duration is an explicit property of each state, along with value [51,54]. Here learning when things will happen proceeds in parallel to learning what will happen, allowing the temporal representation to adapt to the properties of the task [55], in line with recent work suggesting that neurons in the striatum indeed represent time in a task-relevant, adaptive way [56•,57,58].

More generally, the efficiency and correctness of different RL algorithms depends critically on how states and actions are represented, highlighting again the centrality of (environment-)appropriate representations in trial-and-error learning. For instance, an active area of research seeks to extend RL to address scenarios that have hierarchical structure (e.g., navigating a building or cooking a meal), in which tasks can be subdivided into a series of simpler subtasks (with their associated subgoals) to aid learning and simplify action selection [59,60]. Many real-world tasks have such structure, and moreover, different tasks often share subcomponents. Thus encapsulation of the policy that achieves a particular subgoal (e.g., finding the elevator) allows easy transfer to other tasks that involve this same subgoal, making the learning of novel tasks more efficient. As a result, algorithmic issues involved in parsing a task to useful subcomponents interact with a computational-level goal of not only solving the current task, but also acquiring a useful 'toolkit' of policies that can be composed to solve many other tasks. This brings to the fore questions about how a task should be optimally hierarchically decomposed [61], and how to quantify the future benefits of a certain decomposition so they may be weighed against the resulting costs to performance and learning of the current task.

This attempt to divine structure in the learning problem relates more broadly to the necessity of dealing effectively with partially-observable environments where current observations (i.e., perceptual cues) are only probabilistically related to the underlying state that generated them. Optimal RL in partially-observable environments is theoretically possible, although practically intractable [62]. The problem is that the non-Markovian observations are inappropriate as inputs to RL algorithms. Instead, one must compute a Bayesian 'belief state' over the underlying states, which does have the Markov property, but is high dimensional and continuous. In practice, approximate solutions have been suggested. For instance, recent work marrying the principles of 'deep' representations with RL has exceeded human performance in simple Atari games [63•]. This work established the robustness and flexibility of relatively naïve RL methods when learning in the space of abstract, generalized representations.

Finally, nodding to the implementational level is the question of which RL algorithm is implemented in the brain, and the inevitable conclusion that several algorithms may be at play at once. Considering the multiplicity of algorithmic solutions that RL offers, each enjoying different strengths and suffering from its own shortcomings, it may indeed be optimal for learning to proceed via parallel, likely interacting, learning and decision making systems [64,65]. Much recent work has studied the interplay between so-called model-based and model-free modes of RL, capturing and explaining the behavioral consequences of calculating reward predictions using prospection in an internal model of the environment (model-based decision making, also called goal-directed behavior), versus using previously cached values that were learned incrementally through experience (model-free decision making, or habitual behavior) [66,67,68•]. This dual-systems approach is increasingly applied to understanding abnormal behavior as an imbalance in computationally distinct learning systems [39,69,70•]. Similarly inspired by the multiplicity of pathways through the basal ganglia, more mechanistic models of action learning have suggested that action values are a product of parallel, balanced, 'go' (excitatory) and 'no-go' (inhibitory) action selection systems, with recent instantiations of this framework directly implementing model-free RL [71•,72]. Such a composite approach to understanding learning processes underlying rich behavior has been attractive in bridging the algorithmic level of RL with its neural implementation, suggesting a mapping from component algorithms to identifiable and distinct neural structures and circuits, as will be discussed below.

The implementational level: dopamine-dependent learning in the basal ganglia

At the final level of the hierarchy, neuroscientists have had considerable success in mapping functions implied by RL algorithms to neurobiological substrates. Whereas some of the computational and algorithmic questions highlighted above revolved around scaling RL to environments with real-world action and state complexity, the problems at the implementational level arise from the sheer complexity of the neural system, as well as the limitations of different experimental methods.

Much of this work has focused on the basal ganglia, a collection of forebrain nuclei that have long been associated with learning, action selection and decision making. The idea that the basal ganglia implement formal RL algorithms stems from the close correspondence between the responses of midbrain dopamine neurons and the reward prediction-error signal at the heart of RL algorithms [10,73]. The striatum, the input structure of the basal ganglia and a primary target of the widely broadcast dopaminergic neuromodulation, is a prime candidate for learning values and biasing action selection so as to implement RL policies. Indeed, plasticity in corticostriatal synapses is modulated by dopamine signaling [74], in perfect accord with the effect of reward prediction errors on learning in RL algorithms [75].

Despite the seemingly direct correspondence between RL and the function of the basal ganglia, many aspects of this mapping remain open. Primary among these is how are reward prediction errors computed by dopaminergic neurons. The circuit mechanism that compares predictions associated with past and present states (as is necessary for computing temporal-difference prediction errors) is the subject of a number of mechanistic models (e.g., [6,76–78]). Many models posit that this computation derives from the difference in sign of direct and indirect inputs from the striatum to dopamine neurons, however, it remains unclear whether these implementations have the necessary fidelity to the known anatomy [79]. Curiously, few models place the burden of the difference computation on local circuitry in dopaminergic areas [80••,81].

Alongside this question, the perennial doubt of whether dopamine signals correspond to prediction error signals lingers. Here, the question has evolved from initial skepticism [82,83] that has mostly been put to rest [84], to understanding the heterogeneity of dopaminergic responses to both appetitive and aversive stimuli [85–88]. Recent findings illustrating the complexity of what was once thought to be a scalar signal broadcast widely — a last haven of apparent simplicity in brain signaling — suggest a relatively broad mapping between the algorithmic and implementation levels, preserving the spirit, if not the letter, of RL algorithms. However, some implementations of RL actually require multiple parallel prediction errors (e.g., hierarchical RL [89] and successor representation frameworks [90,91]), suggesting that diversity is perhaps a feature of the system rather than a bug. Still, the relationship between dopamine signals and areas that represent positive prediction errors for aversive stimuli (such as the lateral habenula [92,93]) is unclear. In addition, dopamine release has long been associated with action initiation [94,95] and energization [82], suggesting a functional role for dopamine beyond reward prediction-error signaling. Dopamine aside, much work has focused on the interplay between dual excitatory (so-called 'go') and inhibitory ('no-go') pathways through the basal ganglia, and whether this architecture can be mapped onto RL action-selection mechanisms [71•,96,97]. New optogenetic and imaging techniques in mice now allow the basic tenets of these dual-pathway action-selection models to be directly tested, with sometimes surprising results [98–100].

Finally, the almost-exclusive focus of research in the last two decades on implementations of model-free incremental RL in the basal ganglia and in dopaminergic signaling has recently been challenged by demonstrations of model-based signals in the same neural substrates [66]. While these findings may be taken to suggest that the basal ganglia do not implement RL-based learning and action selection, an alternative interpretation is that model-based and model-free RL methods are not as separated in the brain as lesion studies first suggested [101]. Perhaps the strong appeal of a simple dichotomic implementation of two different algorithmic solutions has led researchers to overly-simplified interpretations of data. Ultimately, the intricate nature of a neural implementation of RL is likely due to the variety of interrelated goals that this system must fulfill in complex environments (optimizing reward, transferring learning to new situations, scaling to the current context, etc.), as well as to the fact that different brain areas function together and not in isolation.

One cautionary tale is that a simple model, while extremely powerful at understanding the functions of different neural substrates, is at the same time limited and should not be expected to explain all aspects of the signaling in these areas. The bidirectional interaction between models and neural findings suggests iterative refinement of the models to widen their scope to more phenomena. Progressively more complex experimental designs that are specifically tailored to asking algorithmic (or even computational level) questions are invaluable in this effort — simple designs can elucidate basic principles, but the reach of the conclusions that one can draw from these experiments should not be overextended. Indeed, optogenetics and other targeted neural manipulations now allow powerful tests of models of RL at the level of neural implementation [102••,103•]. However, the ability to transiently perturb neural activity with cell-type selectivity and temporal precision during ongoing behavior demands that precise hypotheses be articulated at the mechanistic level, and behavioral paradigms be sensibly exploited.

Conclusion — inspiration across levels

Reinforcement learning is perhaps the poster child of Marr's levels of analysis — a computational problem that, expressed formally, leads to a host of algorithmic solutions that seem to be implemented in human and animal brains. However, as with many classification schemes, too much emphasis on delineation of levels can distract from the holistic nature of scientific inquiry. As we have shown, the boundaries between the levels are not clear cut, and cross-disciplinary interaction among researchers from different fields and focusing on different levels has only served to advance the field. By some accounts, the different levels should be used to constrain each other — implementational limitations determining which algorithms are feasible, algorithmic (in)efficiencies affecting which computational problems are solvable, etc. However, we feel that this approach is optimistic at best, and misleading at worst — given the degrees of freedom at each level, hard constraints are difficult to derive, and may lead to unnecessary restrictions on creativity at other levels. Instead, to paraphrase Rich Sutton (personal communication, Barbados 2008), Marr's levels serve scientific inquiry much better when used to inspire one another.

Conflict of interests

Nothing declared.

Acknowledgements

We are grateful to Gecia Bravo-Hermsdorff, Mingbo Cai, Andra Geana, Nina Rouhani, Nico Schuck and Yeon Soon Shin for valuable comments on this manuscript. This work was funded by the Human Frontier Science Program Organization and by NIMH grant R01MH098861.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Marr D, Poggio T: **From understanding computation to understanding neural circuitry**. *Neurosci Res Program Bull* 1977, **15**:470-488.

2. Sutton RS, Barto AG: *Reinforcement learning: an introduction*. MIT Press; 1998.

3. Sutton RS: **Learning to predict by the methods of temporal differences**. *Mach Learn* 1988, **3**:9-44.

4. Niv Y, Schoenbaum G: **Dialogues on prediction errors**. *Trends Cogn Sci* 2008, **12**:265-272.

5. Niv Y: **Reinforcement learning in the brain**. *J Math Psychol* 2009, **53**:139-154.

6. Houk JC, Adams JL, Barto AG: **A model of how the basal ganglia generate and use neural signals that predict reinforcement**. In *Models of information processing in the basal ganglia.* Edited by Houk JC, Davis JL, Beiser DG. MIT Press; 1995:249-270.

7. O'Reilly RC, Frank MJ: **Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia**. *Neural Computat* 2006, **18**:283-328.

8. Redgrave P, Prescott TJ, Gurney K: **The basal ganglia: a vertebrate solution to the selection problem?** *Neuroscience* 1999, **89**:1009-1023.

9. Montague PR, Dayan P, Sejnowski TJ: **A framework for mesencephalic dopamine systems based on predictive Hebbian learning**. *J Neurosci* 1996, **16**:1936-1947.

10. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward**. *Science* 1997, **275**:1593-1599.

11. Barto AG: **Adaptive critics and the basal ganglia**. In *Models of information processing in the basal ganglia.* Edited by Houk JC, Davis JL, Beiser DG. MIT Press; 1995:215-232.

12. Kang MJ, Hsu M, Krajbich IM, Loewenstein G, McClure SM, Wang JT, Camerer CF: **The wick in the candle of learning epistemic curiosity activates reward circuitry and enhances memory**. *Psychol Sci* 2009, **20**:963-973.

13. Loewenstein G: **The psychology of curiosity: a review and reinterpretation**. *Psychol Bull* 1994, **116**:75.

14. Schmidhuber J: **A possibility for implementing curiosity and boredom in model-building neural controllers**. *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior*. 1991.

15. Oudeyer P-Y, Kaplan F: **What is intrinsic motivation? A typology of computational approaches**. *Front Neurorobot* 2007, **1**:6.

16. Barto AG: **Intrinsic motivation and reinforcement learning**. In *Intrinsically motivated learning in natural and artificial systems.* Edited by Baldassarre G, Mirolli M. Springer; 2013:17-47.

17. imşek Ö, Barto AG: **An intrinsic reward mechanism for efficient exploration**. In *Proceedings of the 23rd international conference on Machine learning*. ACM; 2006:833-840.

18. Singh S, Lewis RL, Barto AG: **Where do rewards come from**. In *Proceedings of the annual conference of the cognitive science society*. 2009:2601-2606.

19. McDevitt M, Dunn R, Spetch M, Ludvig E: **When good news leads to bad choices**. *J Exp Anal Behav* 2016, **105**:23-40.

20. Pisklak JM, McDevitt MA, Dunn RM, Spetch ML: **When good pigeons make bad decisions: choice with probabilistic delays and outcomes**. *J Exp Anal Behav* 2015, **104**:241-251.

21. Bromberg-Martin ES, Hikosaka O: **Midbrain dopamine neurons signal preference for advance information about upcoming rewards**. *Neuron* 2009, **63**:119-126.

22. Bromberg-Martin ES, Hikosaka O: **Lateral habenula neurons signal errors in the prediction of reward information**. *Nature Neurosci* 2011, **14**:1209-1216.

23. Blanchard TC, Hayden BY, Bromberg-Martin ES: **Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity**. *Neuron* 2015, **85**:602-614.

24. Singh S, Lewis RL, Barto AG, Sorg J: **Intrinsically motivated reinforcement learning: An evolutionary perspective**. *IEEE Trans Autonom Mental Dev* 2010, **2**:70-82.

25. Guo X, Singh S, Lewis RL: **Reward mapping for transfer in long-lived agents**. *Adv Neural Inform Process Syst* 2013:2130-2138.

26. Sorg J, Singh SP, Lewis RL: **Internal rewards mitigate agent boundedness**. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010:1007-1014.

27. Hull C: *Principles of behavior: an introduction to behavior theory*. D. Appleton-Century Company; 1943.

28. • Keramati M, Gutkin B: **Homeostatic reinforcement learning for integrating reward collection and physiological stability**. *eLife* 2014, **3**:e04811.

The authors introduce a new RL theory in which rewards are defined as outcomes that fulfill physiological needs, and use this framework to show how animals learn to perform actions that lead to rewards in order to maintain a stable physiological state. The model can account for a range of phenomena related to food seeking and risk aversion, and suggests a normative understanding of delay discounting as a consequence of optimally minimizing physiological deviations.

29. Keramati M, Gutkin BS: **A reinforcement learning theory for homeostatic regulation**. *Advances in neural information processing systems*. 2011:82-90.

30. Nahum-Shani I, Smith SN, Tewari A, Witkiewitz K, Collins LM, Spring B, Murphy S: **Just in time adaptive interventions (JITAIs): an organizing framework for ongoing health behavior support**. *Methodology Center Technical Report No. 14-126*. 2014.

31. Nahum-Shani I, Hekler EB, Spruijt-Metz D: **Building health behavior models to guide the development of just-in-time adaptive interventions: a pragmatic framework**. *Health Psychol* 2015, **34**:1209.

32. Konidaris G, Scheidwasser I, Barto AG: **Transfer in reinforcement learning via shared features**. *J Mach Learn Res* 2012, **13**:1333-1371.

33. Gershman SJ, Niv Y: **Learning latent structure: carving nature at its joints**. *Curr Opin Neurobiol* 2010, **20**:251-256.

34. Daw N: **Trial by trial data analysis using computational models.Decision making, affect and learning: attention and performance XXIII**, Oxford University Press; 2011.

35. Gershman SJ, Norman KA, Niv Y: **Discovering latent causes in reinforcement learning**. *Curr Opin Behav Sci* 2015, **5**:43-50.

36. Gershman SJ, Blei DM, Niv Y: **Context, learning, and extinction**. *Psychol Rev* 2010, **117**:197.

37. Gershman SJ, Jones CE, Norman KA, Monfils M-H, Niv Y: **Gradual extinction prevents the return of fear: implications for the discovery of state**. *Front Behav Neurosci* 2013, **7**:164.

38. Gershman SJ, Niv Y: **Perceptual estimation obeys Occam's razor**. *Front Psychol* 2013, **4**:623.

39. Collins AG, Brown JK, Gold JM, Waltz JA, Frank MJ: **Working memory contributions to reinforcement learning impairments in schizophrenia**. *J Neurosci* 2014, **34**:13747-13756.

40. Gershman SJ, Radulescu A, Norman KA, Niv Y: **Statistical computations underlying the dynamics of memory updating**. *PLoS Computat Biol* 2014, **10**:e1003939.

41. Niv Y, Daniel R, Geana A, Gershman SJ, Leong YC, Radulescu A, Wilson RC: **Reinforcement learning in multidimensional environments relies on attention mechanisms**. *J Neurosci* 2015, **35**:8145-8157.

42. Glimcher PW, Fehr E: *Neuroeconomics: decision making and the brain*. Academic Press; 2013.

43. Levy DJ, Glimcher PW: **The root of all value: a neural common currency for choice**. *Curr Opin Neurobiol* 2012, **22**:1027-1038.

44. Padoa-Schioppa C, Assad JA: **Neurons in the orbitofrontal cortex encode economic value**. *Nature* 2006, **441**:223-226.

45. Padoa-Schioppa C, Assad JA: **The representation of economic value in the orbitofrontal cortex is invariant for changes of menu**. *Nature Neurosci* 2008, **11**:95-102.

46. Baxter J, Bartlett PL: **Infinite-horizon policy-gradient estimation**. *J Artif Intell Res* 2001:319-350.

47. Sutton RS, McAllester DA, Singh SP, Mansour Y: **Policy gradient methods for reinforcement learning with function approximation**. *Adv Neural Inform Process Syst* 1999:1057-1063.

48. Bhatnagar S, Ghavamzadeh M, Lee M, Sutton RS: **Incremental natural actor-critic algorithms**. *Adv Neural Inform Process Syst* 2007:105-112.

49. Li J, Daw ND: **Signals in human striatum are appropriate for policy update rather than value prediction**. *J Neurosci* 2011, **31**:5504-5511.

50. Gershman SJ, Moustafa AA, Ludvig EA: **Time representation in reinforcement learning models of the basal ganglia**. *Front Computat Neurosci* 2014:7.

51. Daw ND, Courville AC, Touretzky DS: **Representation and timing in theories of the dopamine system**. *Neural Computat* 2006, **18**:1637-1677.

52. Ludvig EA, Sutton RS, Kehoe EJ: **Evaluating the TD model of classical conditioning**. *Learn Behav* 2012, **40**:305-319.

53. Rivest F, Kalaska JF, Bengio Y: **Alternative time representation in dopamine models**. *J Computat Neurosci* 2010, **28**:107-130.

54. Bradtke SJ, Duff MO: **Reinforcement learning methods for continuous time Markov decision problems**. *Adv Neural Inform Process Syst* 1995, **7**:393.

55. Takahashi Y, Langdon AJ, Niv Y, Schoenbaum G: **Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum**. *Neuron* 2016.

56. Mello GB, Soares S, Paton JJ: **A scalable population code for time in the striatum**. *Curr Biol* 2015, **25**:1113-1122.
•
Using a serial fixed interval task, the authors find a population of neurons in the dorsal striatum that adaptively represents time. An adaptive temporal representation such as this is one of the essential components of RL in semi-Markov environments.

57. Gouvêa TS, Monteiro T, Motiwala A, Soares S, Machens C, Paton JJ: **Striatal dynamics explain duration judgments**. *eLife* 2016, **4**:e11386.

58. Jin X, Tecuapetla F, Costa RM: **Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences**. *Nature Neurosci* 2014, **17**:423-430.

59. Botvinick MM, Niv Y, Barto AC: **Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective**. *Cognition* 2009, **113**:262-280.

60. Botvinick MM: **Hierarchical reinforcement learning and decision making**. *Curr Opin Neurobiol* 2012, **22**:956-962.

61. Solway A, Diuk C, Córdova N, Yee D, Barto AG, Niv Y, Botvinick MM: **Optimal behavioral hierarchy**. *PLoS Computat Biol* 2014, **10** e1003779.

62. Kaelbling LP, Littman ML, Cassandra AR: **Planning and acting in partially observable stochastic domains**. *Artif Intell* 1998, **101**:99-134.

63. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G: **Human-level control through deep reinforcement learning**. *Nature* 2015, **518**:529-533.
•
The authors marry a deep convolutional artifical neural network with the Q-learning algorithm from RL to produce an agent (called the deep Q-network; DQN) that learns to play a range of Atari games at human level or beyond. With only the game pixel display and the score as input, the DQN agent is able to learn a generalized, high-dimensional action-value representation of the display to support flexible game-playing behavior.

64. Daw ND, Niv Y, Dayan P: **Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control**. *Nature Neurosci* 2005, **8**:1704-1711.

65. Keramati M, Dezfouli A, Piray P: **Speed/accuracy trade-off between the habitual and the goal-directed processes**. *PLoS Computat Biol* 2011, **7** e1002055.

66. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ: **Model-based influences on humans' choices and striatal prediction errors**. *Neuron* 2011, **69**:1204-1215.

67. Otto AR, Gershman SJ, Markman AB, Daw ND: **The curse of planning dissecting multiple reinforcement-learning systems by taxing the central executive**. *Psychol Sci* 2013, **24**:751-761.

68. Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND: **Model-based choices involve prospective neural activity**. *Nature Neurosci* 2015, **18**:767-772.
•
Behavioral work has suggested that planning in model-based RL involves prospection at decision time. Neural findings of hippocampal preplay offer a neural substrate for this, but have not been connected to choices. Here Doll *et al*. use a clever task in humans undergoing fMRI to show that the extent to which participants behave according to model-based RL is correlated with the decodability of prospective representations in their brain, and conversely, prediction error signals can be detected in participants' striatum to the degree that they are using model-free RL.

69. Voon V, Derbyshire K, Rück C, Irvine M, Worbe Y, Enander J, Schreiber L, Gillan C, Fineberg N, Sahakian B: **Disorders of compulsivity: a common bias towards learning habits**. *Mol Psychiatry* 2015, **20**:345-352.

70. Gillan C, Kosinski RW, Phelps EA, Daw ND: **Characterizing a psychological dimension related to deficits in goal-directed control**. *eLife* 2016, **5**:e11305.
•
Using a large sample of thousands of healthy participants on Amazon's Mechanical Turk, the authors identify compulsivity, anxious depression and social withdrawal as factors that cut across symptoms of different mental disorders. They then use a sequential choice task to quantify the relative contribution of model-based versus model-free RL to each participant's decisions, and show that reduced model-based control is specifically related to increased compulsivity.

71. Collins AG, Frank MJ: **Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive**. *Psychol Rev* 2014, **121**:337.
•
The authors present an expanded actor-critic RL algorithm that incorporates symmetric and balanced positive and negative action-value learning to reflect the parallel 'go' and 'no-go' pathways found in the basal ganglia. This model accounts for seemingly incompatible effects of dopamine on learning and on motivation. The authors go on to show that such an apparently redundant action-representation mechanism might be normative, as it allows the agent to learn equally well in 'rich' as in 'lean' environments.

72. Cockburn J, Collins AG, Frank MJ: **A reinforcement learning mechanism responsible for the valuation of free choice**. *Neuron* 2014, **83**:551-557.

73. Roesch MR, Calu DJ, Schoenbaum G: **Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards**. *Nature Neurosci* 2007, **10**:1615-1624.

74. Reynolds JN, Hyland BI, Wickens JR: **A cellular mechanism of reward-related learning**. *Nature* 2001, **413**:67-70.

75. Reynolds JN, Wickens JR: **Dopamine-dependent plasticity of corticostriatal synapses**. *Neural Netw* 2002, **15**:507-521.

76. Morita K, Morishima M, Sakai K, Kawaguchi Y: **Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways**. *Trends Neurosci* 2012, **35**:457-467.

77. Potjans W, Diesmann M, Morrison A: **An imperfect dopaminergic error signal can drive temporal-difference learning**. *PLoS Computat Biol* 2011, **7** e1001133.

78. Potjans W, Morrison A, Diesmann M: **A spiking neural network model of an actor-critic learning agent**. *Neural Computat* 2009, **21**:301-339.

79. Joel D, Niv Y, Ruppin E: **Actor–critic models of the basal ganglia: new anatomical and computational perspectives**. *Neural Netw* 2002, **15**:535-547.

80. •• Eshel N, Bukwich M, Rao V, Hemmelder V, Tian J, Uchida N: **Arithmetic and local circuitry underlying dopamine prediction errors**. *Nature* 2015, **525**:243-246.

Using optogenetics and extracellular recordings, the authors demonstrate that dopamine neurons in the ventral tegmental area (VTA) perform subtraction, and that this operation depends critically on the activity of inhibitory interneurons within the VTA. This is the first demonstration in this circuit of this arithmetic operation, central to the computation of reward prediction errors in RL.

81. Rao RP, Sejnowski TJ: **Spike-timing-dependent Hebbian plasticity as temporal difference learning**. *Neural Computat* 2001, **13**:2221-2237.

82. Berridge KC: **The debate over dopamine's role in reward: the case for incentive salience**. *Psychopharmacology* 2007, **191**:391-431.

83. Redgrave P, Prescott TJ, Gurney K: **Is the short-latency dopamine response too short to signal reward error?** *Trends Neurosci* 1999, **22**:146-151.

84. Schulz JM, Redgrave P, Mehring C, Aertsen A, Clements KM, Wickens JR, Reynolds JN: **Short-latency activation of striatal spiny neurons via subcortical visual pathways**. *J Neurosci* 2009, **29**:6336-6347.

85. Cohen JY, Haesler S, Vong L, Lowell BB, Uchida N: **Neuron-type-specific signals for reward and punishment in the ventral tegmental area**. *Nature* 2012, **482**:85-88.

86. Henny P, Brown MT, Northrop A, Faunes M, Ungless MA, Magill PJ, Bolam JP: **Structural correlates of heterogeneous in vivo activity of midbrain dopaminergic neurons**. *Nature Neurosci* 2012, **15**:613-619.

87. 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**:4894-4899.

88. Joshua M, Adler A, Mitelman R, Vaadia E, Bergman H: **Midbrain dopaminergic neurons and striatal cholinergic interneurons encode the difference between reward and aversive events at different epochs of probabilistic classical conditioning trials**. *J Neurosci* 2008, **28**:11673-11684.

89. Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y: **Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia**. *J Neurosci* 2013, **33**:5797-5805.

90. Gershman SJ, Moore CD, Todd MT, Norman KA, Sederberg PB: **The successor representation and temporal context**. *Neural Computat* 2012, **24**:1553-1568.

91. Dayan P: **Improving generalization for temporal difference learning: the successor representation**. *Neural Computat* 1993, **5**:613-624.

92. Matsumoto M, Hikosaka O: **Lateral habenula as a source of negative reward signals in dopamine neurons**. *Nature* 2007, **447**:1111-1115.

93. Matsumoto M, Hikosaka O: **Representation of negative motivational value in the primate lateral habenula**. *Nature Neurosci* 2009, **12**:77-84.

94. Syed EC, Grima LL, Magill PJ, Bogacz R, Brown P, Walton ME: **Action initiation shapes mesolimbic dopamine encoding of future rewards**. *Nature Neurosci* 2016, **19**:34-36.

95. Phillips PE, Stuber GD, Heien ML, Wightman RM, Carelli RM: **Subsecond dopamine release promotes cocaine seeking**. *Nature* 2003, **422**:614-618.

96. Frank MJ, Seeberger LC, O'Reilly RC: **By carrot or by stick: cognitive reinforcement learning in parkinsonism**. *Science* 2004, **306**:1940-1943.

97. Cox SM, Frank MJ, Larcher K, Fellows LK, Clark CA, Leyton M, Dagher A: **Striatal D1 and D2 signaling differentially predict learning from positive and negative outcomes**. *Neuroimage* 2015, **109**:95-101.

98. Tai L-H, Lee AM, Benavidez N, Bonci A, Wilbrecht L: **Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value**. *Nature Neurosci* 2012, **15**:1281-1289.

99. Cui G, Jun SB, Jin X, Pham MD, Vogel SS, Lovinger DM, Costa RM: **Concurrent activation of striatal direct and indirect pathways during action initiation**. *Nature* 2013, **494**:238-242.

100. Calabresi P, Picconi B, Tozzi A, Ghiglieri V, Di Filippo M: **Direct and indirect pathways of basal ganglia: a critical reappraisal**. *Nature Neurosci* 2014, **17**:1022-1030.

101. Balleine BW: **Neural bases of food-seeking: affect, arousal and reward in corticostriatolimbic circuits**. *Physiol Behav* 2005, **86**:717-730.

102. •• Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH: **A causal link between prediction errors, dopamine neurons and learning**. *Nature Neurosci* 2013, **16**:966-973.

The authors use the behavioral paradigm of blocking together with optogenetic manipulation of dopamine neurons to show that dopamine activity is sufficient to generate learning, as predicted by RL theories.

103. • Chang CY, Esber GR, Marrero-Garcia Y, Yau H-J, Bonci A, Schoenbaum G: **Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors**. *Nature Neurosci* 2016, **19**:111-116.

The authors use optogenetics to briefly suppress firing activity of dopamine neurons and show that this manipulation alters learning in a Pavlovian overexpectation task, consistent with the effect of a negative prediction error in RL.