Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Reinforcement-guided learning in frontal neocortex: emerging computational concepts

Abhishek Banerjee[1], Rajeev V Rikhye[2] and Adam Marblestone[3]

The classical concepts of reinforcement learning in the mammalian brain focus on dopamine release in the basal ganglia as the neural substrate of reward prediction errors, which drive plasticity in striatal and cortico-striatal synapses to maximize the expected aggregate future reward. This temporal difference framework, however, even when augmented with deep credit assignment, does not fully capture higher-order processes such as the influence of goal representations, planning based on learned internal models, and hierarchical decision-making implemented by diverse neocortical areas. Candidate functions for such neocortical contributions to reinforcement learning are increasingly being considered in artificial intelligence algorithms. Here, we review recent experimental neurophysiological findings focusing on the orbitofrontal cortex, a key higher-order association cortex, and highlight emerging concepts that emphasize the role of the neocortex in reward-driven computation, in addition to its role as an input to striatal structures. In this framework, reward drives plasticity in various neocortical regions, implementing multiple distinct reinforcement learning algorithms.

Addresses
[1] Neurosciences Theme, Biosciences Institute, Newcastle University, United Kingdom
[2] MIT Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States
[3] Federation of American Scientists, United States

Corresponding author:
Banerjee, Abhishek (abhi.banerjee@newcastle.ac.uk)

Current Opinion in Behavioral Sciences 2021, **38**:133–140

This review comes from a themed issue on **Computational cognitive neuroscience**

Edited by **Geoff Schoenbaum** and **Angela J Langdon**

https://doi.org/10.1016/j.cobeha.2021.02.019

Introduction

Reinforcement learning (RL) models have been useful for understanding learning in biological systems including the mammalian brain, as well as for building artificial intelligence agents, which has achieved super-human like performance on a variety of tasks [1]. In studies of the mammalian brain, theoretical and computational models of RL have primarily focused on the role of subcortical structures, such as the ventral tegmental area (VTA) and the basal ganglia, within the concept of dopamine-based RL [2–4]. In such models, reward prediction errors (RPE; see **Glossary**) are computed subcortically which then drive adaptive plasticity predominantly in the striatal and cortico-striatal synapses. Recordings and causal manipulations of mouse VTA dopamine (DA) neurons confirm this basic concept by demonstrating that these neurons use 'subtraction' as arithmetic operation to calculate RPE signals, implemented by neighboring GABAergic neurons, and finally broadcast especially to the striatal neurons [5•].

Moving beyond these classical models, an emerging literature now explores complex reward-related learning in the neocortex, suggesting that actions and rewards drive further plasticity within a complex network of neocortical areas [6••,7–9]. These areas themselves represent several reward-related characteristics, including reward history and utility, value predictions, expectation errors, and accumulation of sensory evidence. Task learning induced changes in such a distributed cortical network presumably are essential for flexible decision-making. Here, we review and discuss recent experimental evidence for complex neocortical contributions to reward processing and suggest approaches to further integrate this framework into computational concepts of RL and their possible mapping onto cortical circuits.

The classical model of reinforcement learning

A classical functional breakdown of the brain's learning systems was provided in the 1990's by Kenji Doya, who attributed unsupervised learning to the neocortex, supervised learning to the cerebellum, and reinforcement learning to the striatum [10]. In the model-free RL (see **Glossary**), the dorsal striatum acts as an 'actor', selecting actions with the largest predicted long-term value given the current state [11–14]. In contrast, the ventral striatum is thought to act as a 'critic', learning a 'value-function' (see **Glossary**) that predicts the aggregate future reward of the animal [15,16]. Dopamine signals from the critic are sent back to the actor, for example, via the VTA, gating the plasticity of synapses onto the medium spiny neurons at the input of the dorsal striatum. Computationally, the challenge is to learn the critic, and this can be achieved using simple 'model-free' RL algorithms such as 'temporal-difference' learning (TD learning), which compares value-functions evaluated at different time steps to drive recursive updating of an estimate

Glossary

Model-free and model-based reinforcement learning (RL): Model-free reinforcement learning relies on a cached or pre-computed state-value mapping, whereas model-based reinforcement learning assumes the ability to engage in planning at task time to roll out potential future consequences of actions.

Meta-learning: Meta-learning, or "learning to learn", is an emerging concept in artificial intelligence research wherein a learning model is optimized not just for performance on single tasks, but for acquisition of the ability to rapidly learn new task variants.

Distributional RL: In environments where rewards and state transitions are inherently stochastic, brain uses this set of algorithms to represent rewards as a probability distribution, effectively representing multiple behaviorally relevant future outcomes.

Value-function: An estimate of the total discounted future reward given the animal's current state.

Reward prediction error (RPE): A quantity used in temporal difference learning models to iteratively update the value function based on previous value function estimates and the instantaneous reward delivered to the animal.

Orbitofrontal cortex (OFC): A key area of the prefrontal cortex in rodents, non-human primates, and humans, that plays a significant role in the flexible control of behaviour and value-based decision-making. OFC is crucially involved in cortical contributions to RL, such as in maintaining goal representations that transiently associate aspects of the animal's relationship with the environment with appetitive or aversive value.

of the discounted total future reward. There is extensive evidence for TD-like learning of dopamine responses in the VTA and other striatal areas [17]. Recent studies on distributional RL additionally suggest that VTA DA neurons may represent possible future rewards not as a single mean scalar quantity, but instead as a probability distribution, effectively representing multiple future outcomes [18•]. Combined with hippocampal-inspired replay of past episodes, model-free RL can learn to perform complex algorithms [1]. Model-free RL in the striatum is an evolutionarily ancient system [19]. While the diversity of DA neuronal responses in the VTA is complex and its actions are still being investigated, the relationship between the basic RL and its neuronal substrates have received much rigorous mechanistic confirmation in rodents, as well as in songbird [20] and other mammals [14,17,21].

In the classical model of RL, the neocortex is often taken to provide an up to date 'state' representation (both external and internal) of task-related information, perhaps learned via an unsupervised representation learning process. The perspective of cortex as a passive provider of state representations needs to be broadened, however, when one incorporates 'model-based' RL (see **Glossary**) requiring forward planning as well as novel goal-driven and hierarchical forms of RL. It is in this context that recent experimental studies on the role of the neocortex in RL shed light on more sophisticated algorithms that go beyond the model-free RL paradigm.

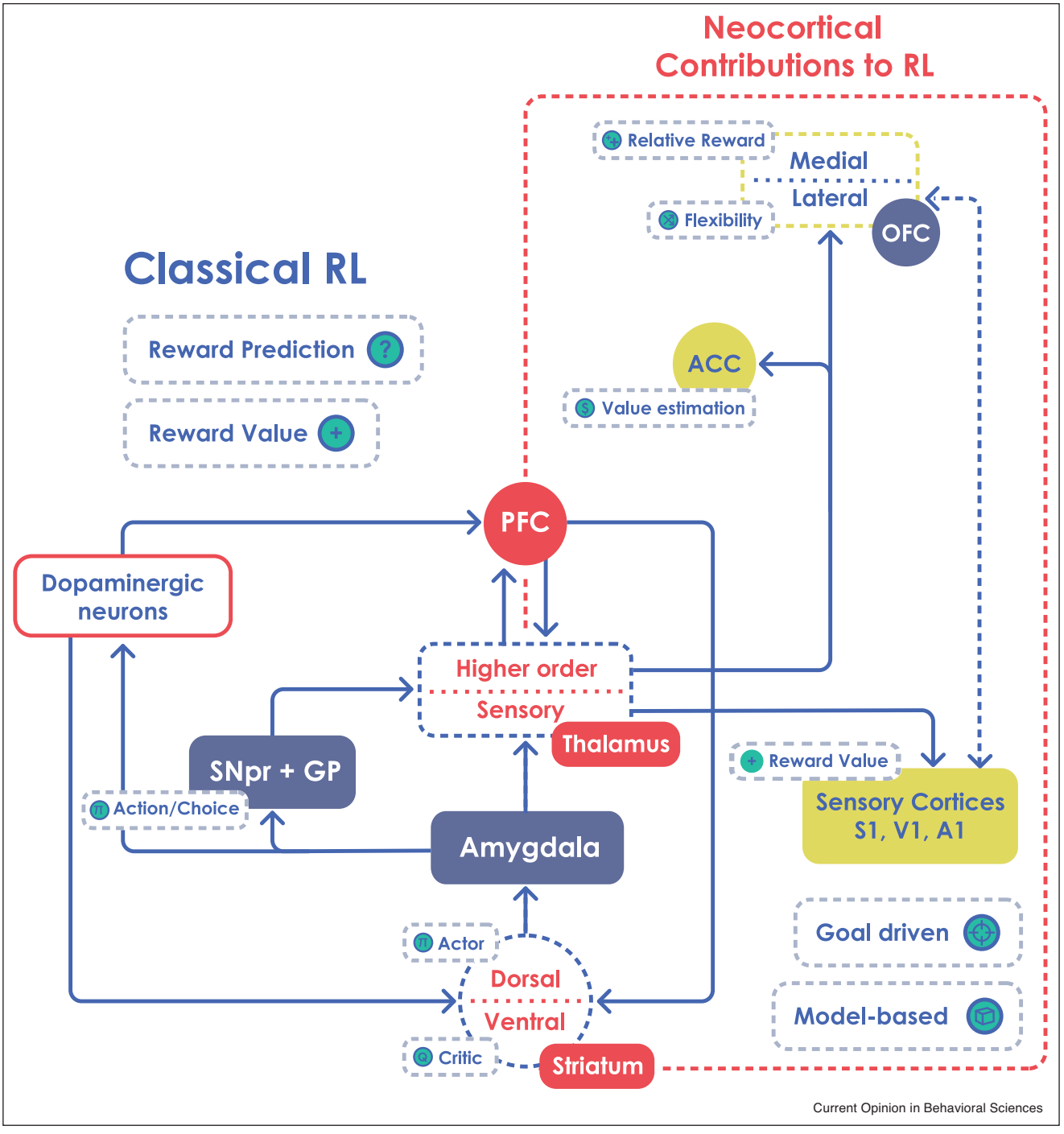
Neocortical contributions to reinforcement learning: neurobiological and computational perspectives

Mammalian brains have several reward-related systems that extend beyond the striatum [22]. Higher-order neocortical areas continuously generate and update predictions of sensory inputs and function as an internal reference framework to compare predicted and actual rewards [9,23] depending on the current context of task [24]. Notably, reward-timing and value information is conveyed even to the primary and associational sensory areas of neocortex [25], including visual [26,27,28••], somatosensory [29,30••], auditory [31], and motor cortices [32]. Converging evidence indicates the involvement of key cortical structures such as the medial prefrontal cortex (mPFC), prelimbic area (PrL), anterior cingulate cortex (ACC), and the orbitofrontal cortex (OFC, see **Glossary**). Together, these areas promote complex goal-directed learning, monitoring task-performance, value-based decision-making, and credit assignment (Figure 1) [6••,33,34,35••]. It is easy to draw parallels to the Markov Decision Process formulation of RL, where different cortical areas encode the state space, the action space, and the probability of that an action in one state will result in a transition to another state at a later time. Task-dependent network dynamics represented in these cortical areas is flexible and can switch and shift between distinct activity states [36]. Studies investigating complex and dynamic interactions between mPFC and striatal circuits governing hedonic responses reveal that elevated excitability in mPFC reduced striatal responses to the stimulation of dopaminergic neurons and modulated reward-seeking behavioral drive [37••] highlighting dynamic long-range cortico-subcortical interactions between RL parameters. More generally, the cortical network, together with cortico-subcortical interplay [35••] involving basolateral amygdala (bLA), and mediodorsal (MD) thalamus, could support more complex forms of RL in several ways beyond just representing states: by computing and updating goal-dependent value-functions, by enabling RL to cope with hierarchical decision-making, by enabling model-based planning to be learned using RL, and potentially by learning to implement entire RL algorithms cortically.

Goal-driven learning

In humans, a growing body of work has emphasized the importance of goal representations in the brain [6••,8]. Humans appear to not simply respond to a monolithic reward, but to engage specific goals at any given time, such as 'find food' versus 'find shelter'. But how can we ground this psychological notion of goals in the brain? One proposal is that goals are 'state-value associations' that become progressively more 'satisfied' as a specific goal is nearer to being achieved, as indicated by associated cues or inferences [6••]. For instance, a goal to find food sets up, effectively, a value-function that progressively becomes higher as the animal gets closer to previous prey locations,

Figure 1



Classical reinforcement learning model and neocortical contributions. In the standard model of reinforcement learning (RL), the medium spiny neurons in the dorsal striatum acts as an 'actor', selecting actions with largest predicted reward value, whereas, the ventral striatum is thought to act as a 'critic'. Outputs of the striatum controls function in several key sub-cortical areas (including the thalamus, superior colliculus, and amygdala, among several others), and to cortical areas (such as the prefrontal and premotor areas) for motor planning and movement preparation, where they may also contribute to decision-making processes. An emerging view accommodates neocortical contributions to goal representations and goal-based value functions serving as a substrate to perform its own context-dependent RL algorithm. Neocortical structures like ACC and OFC, and even sensory cortices (S1, V1, and A1) plays an important role in computing complex goal representations including reward timing and value predictions, estimates of confidence and behavioural flexibility.

sees the prey, approaches, and so forth. Overall, having goals makes decision-making flexible, adaptive and context-dependent, and supports model-based planning (see Ref. [38] for a detailed overview).

Goal-driven learning and the OFC

A network of frontal cortical regions supports goal-driven aspects of RL to represent goals, monitor and progress towards them, as well as to mediate switching between goals. Distinct frontal cortical regions show a remarkable and diverse set of behavioral and cognitive functions [39]. One of the key brain areas in higher mammals that has been implicated in such processes is the OFC [40,41]. OFC function is critical for flexible stimulus-reward learning, and for behavior based on expected outcome values that require integrating the sensory features of potential choice outcomes. Human neuroimaging studies and single-unit electrophysiological studies in non-human primate (NHP) have confirmed a role for OFC in coding stimulus-value from a variety of sensory modalities, including olfaction, somatosensation, audition, and vision, linking reward to hedonic experiences [40]. Recent human fMRI studies have also shown that error signals in the midbrain are conveyed to the OFC and other cortical target areas [42]. O'Reilly *et al.* posit "a rich semantic map of a very high-dimensional goal space distributed across ventral/medial PFC (OFC and ACC) and other areas of cortex", describing a hierarchy of goal-related areas [6••]. At the lowest levels of the proposed hierarchy are areas that respond to 'primary values', such as the direct receipt of food or physical comfort, for example, the hypothalamus. At one level higher, amygdala areas associate learned cues with these primary values, for example, recognizing that the smell of honey is associated with both food and bee stings. Further up, areas such as amygdala and the ventral striatum develop a more sophisticated and predictive value-based context-dependent functions and modulate the lower areas. At the top of the hierarchy, OFC maintains high-level goal representations through persistent activity even in the absence of a stimulus [43]. Further areas like the ACC interact with high-level action plan representations (in the dorsolateral PFC and other PFC areas) to associate goals with plans for achieving them and the costs, conflicts, uncertainties and tradeoffs involved.

Physiological and lesion studies are consistent with this functional view of OFC in value-updating when assigning relative value and identity to two alternative goals [41,44]. Rodents show significant impairment in reversal learning upon pharmacogenetic silencing of OFC [30••,45]. Similar silencing using optogenetic tools prevented experience-based updating of outcome evaluation [46]. There is strong evidence that dynamic interactions between OFC and dorsolateral striatum critically control the balance between goal-directed versus habitual actions and OFC conveys information about action values [47]. In addition, these interactions support real-time cognitive utilization of reward estimations [48].

Learning engages OFC circuits across sensory modalities [30••,49]. Upon learning a sensory discrimination task, a large fraction of OFC neurons responds to rewarded conditioned stimuli as well as internal state and task context and outcome [30••,49]. Detailed functional analysis has shown that OFC neurons use prior knowledge to facilitate learning in an odor sequence task and the neuronal ensembles converge onto a low-dimensional code [50••]. Neuronal activity in the OFC is proposed to be causal to economic choices. High-current stimulation of OFC was found to disrupt the comparison of subjective values in NHP and in turn found to increase choice variability [51]. Electrophysiological studies in the NHP highlight that OFC neurons encode value in phase with theta oscillation and disrupting theta with micro stimulation impairs learning of new values [52•].

Orbitofrontal prediction broadcast signals

Rodent studies have uncovered a crucial role of OFC neurons in globally broadcasting reward and RPE-related signals to impact other areas. OFC circuits contribute to multiple RL processes by distinctly involving specific projections controlling components of the action-value updating [53], similar to frontal pre-motor cortex's role in cortical representation of task engagement [54•]. By performing high-density electrocorticogram recordings in NHP solving an auditory task, another study similarly found prediction and prediction-error signals being encoded in the PFC that are dynamically conveyed to temporal cortex [55•]. This functional framework makes predictions for the interaction of OFC with other cortical and subcortical areas such as the striatum. O'Reilly *et al.* contrast the role of dorsal striatum in action selection with the role of ventral striatum in goal selection [6••]. In their model, ventral striatum controls how goal representations that are held in working memory are transferred to and from the OFC and other prefrontal areas. Consistent with this idea, evidence has been found for ventral striatum gating OFC representations in mice [56]. Interestingly, a recent rodent neurophysiology and optogenetics study found a role for OFC value representations in learning but not directly in choice [57]. Value computation could also importantly determine abstractions by directly affecting early sensory areas through top-down modulation [30••]. In a reversal learning task based on tactile discrimination, OFC neurons were found to encode both 'positive' and 'negative' prediction errors showing increased responses for new 'hit' as well as 'false alarm' after rule-switch [30••]. This RPE signal implemented functional remapping in the sensory cortex via direct lateral OFC projections carrying RPE information [30••], as well as via gain modulation of irrelevant stimuli [58].

Computational implications of neocortical goal-driven reinforcement learning

What implications might goal-driven reinforcement learning have for the computational frameworks used to model such experiments across species? In a computational model that uses goals, Solway and Botvinick focused on the process of hierarchical planning to achieve goals [8]. In this model, Bayesian computations to invert a generative model of reward are proposed to be performed by a network including 1) dorsolateral PFC for contingency-based or rule-based action plan representations, 2) medial temporal lobe/hippocampus as well as striatum for predicting action consequences or future states, and 3) OFC and bLA for predicting reward contingencies for modeled outcomes. This model differs significantly from the value-function approximation and policy search methods prominent in RL-based AI today. A few other neural network AI models have started to bridge to goal-based frameworks, such as 'universal value-function approximators' [59], in which value-function approximation is made conditional on goals. Despite these studies and models, a major gap still exists between goal-driven views of functional anatomy from cognitive science (with their emerging and still quite tentative basis of empirical support from detailed neurophysiology in rodents and other small animals [60]) and modern high-performance RL algorithms.

Other emerging computational frameworks

We have focused on goal-driven RL and the role of OFC here, but more broadly, cortical contributions to RL may be crucial in other aspects. For example, in *hierarchical reinforcement learning* [61], elementary actions are chunked into groups of macro-actions such that planning and learning can occur at the macro-action level. In cognitive neuroscience, human imaging has pinpointed a signature of hierarchical RL across areas including the nucleus accumbens, ACC, habenula, and amygdala [62]. Rodent studies have suggested that the action space at the output of the basal ganglia is itself hierarchically structured [63], while other studies suggest that the basal ganglia can learn sequences [64]. In computational cognitive modeling, Frank *et al.* have proposed a PFC- and basal ganglia-dependent network model of 'strategic cognitive sequencing' for hierarchical control and learning [65], based on attractor-like constraint satisfaction processing to associate goals and present states with useful sub-goals.

Additionally, model based RL has long been suggested to rely on key cortical contributions that intersect with neural substrates of planning and rule-based cognition. The model based versus model free distinction is extensively studied in cognitive science, where a role for dorsolateral PFC in model-based planning has been suggested [14], a role of the striatum in model-free RL based control, and an arbitration between the systems depending on the uncertainty level. Also, successor representations have been suggested to form a kind of intermediate point on a spectrum between model-based and model-free RL, which might facilitate sub-goal discovery for hierarchical planning and explain coding properties of the hippocampus in rodent navigation [66,67]. The successor representation makes it computationally easy to rapidly update state-value associations but relies on a slower process to learn predictive state representations. While there have been several parallels between successor representations and the hippocampus [67], there are several limitations to this representation that make it intractable for the cortex. First, by aggregating value functions over time, the successor representation does not allow the agent to preserve the temporal ordering of task states. Second, the successor representation is linked to each policy and have to be recomputed each time the reward changes. The cortex however should be able to update policies on-the-fly. Given these drawbacks, the notion that the OFC represents a cognitive map of task space is more apt for cortical RL [68,69]. In this hypothesis, OFC represents a cognitive map of the task space of unobservable information, such as working memory. This hypothesis successfully recapitulates many of the observed roles of the OFC including its role in reversal learning and credit assignment amongst others. Importantly this cognitive map hypothesis places the OFC in the center of both model-free and model-based RL. Specifically, using this cognitive map, agents can take advantage of learned relationship between states in order to vicariously plan their next course of action. Computationally, a cognitive map or a graph, is an efficient means of storing goal (or location, in the case of the hippocampus) information. A machine learning study using probabilistic graphical models recently showed that, through random exploration and without explicit RL algorithms, an agent could learn a cognitive graph of its environment [70]. Using this map, the agent was able to find efficient routes within a complicated environment and plan hierarchically. We believe that the OFC cognitive map could function in a similar way, allowing intelligent agents to efficiently encode goals representations and plan between task states. This theme could be common among the PFC, which represents rule information, and the hippocampus, which represents spatial contextual information, via a cognitive graph [70].

Finally, PFC has been proposed to play a key role as RL system in itself, in a *meta-learning* setting where the slow, biologically inbuilt striatal dopaminergic system effectively trains the PFC to itself carry out an RL algorithm through its fast activity dynamics [71••]. The model could explain several experimental findings, including the human adaptation of RL learning rates to the volatility of reward, and primate findings in which the RL system is able to infer values of states that have not been experienced. The model also predicts that PFC would encode many features of RL algorithms such as learning rates, value functions, RPE and so forth. This possibility of full RL algorithms emerging inside cortical networks further complicates the interpretation of cortical contributions to RL and their delineation from striatal contributions. Interestingly, this entire system was trained using model-free RL, yet it gave rise to a more sophisticated

learning algorithm through a process of 'meta-learning' across many tasks. Thus, some of the signatures of RL in cortex may be dedicated learning mechanisms built by evolution, whereas others may be emergent from model-free learning and credit assignment in the context of a rich and complex (e.g. multi-task) environment.

Conclusion

In conclusion, the neocortex could contribute to RL in several ways - including computing goal representations and goal-based value functions, supporting hierarchical and model-based learning, computing reward history involving specific neuronal ensembles, and implementing its own context-dependent RL algorithm. In each case, the concepts arise from a mixture of human and NHP behavioral and imaging studies on the one side, and computational models from both computational cognitive science [72] and AI [73] on the other side. These fields are now beginning to make contact with detailed neurophysiological studies in rodents, as well as purely behavioral studies revealing effects like cost sensitivity [74••]. Possible new mechanistic studies include combinations of optogenetic perturbations with activity measurements either globally, in specific pathways, or in specific cell types. Combining advanced experimental methods in rodents with more complex yet ethological tasks that require learning hierarchical and context-dependent decision rules should serve to further link mechanistic circuit analysis with cognitive and computational frameworks and reveal the roles and interactions between specific mesoscale cortical and sub-cortical areas and pathways in the context of an emerging computational understanding of the possibilities afforded by multi-component RL algorithms.

Conflict of interest statement

Nothing declared.

Acknowledgements

This work is supported by a Royal Society research grant (to A.B.). We thank A. Marques-Smith and Blake Richards for helpful discussions and D. Haydock in reading the manuscript. The authors thank Drs. Fritjof Helmchen and Christopher Lewis for their comments on an earlier version of the manuscript.

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. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G *et al.*: **Human-level control through deep reinforcement learning**. *Nature* 2015, **518**:529-533.

2. Glimcher PW: **Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis**. *Proc Natl Acad Sci U S A* 2011, **108**:15647-15654.

3. Watabe-Uchida M, Eshel N, Uchida N: **Neural circuitry of reward prediction error**. *Annu Rev Neurosci* 2017, **40**:373-394.

4. Bush RR, Mosteller F: **A mathematical model for simple learning**. *Psychol Rev* 1951, **58**:313-323.

5. • 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

A mechanistic investigation of how RPE are computed in the mammalian brain, demonstrating the use of a subtraction computation that fits naturally with computational models of RPE computation.

6. •• O'Reilly RC, Hazy TE, Mollick J, Mackie P, Herd S: *Goal-Driven Cognition in the Brain: A Computational Framework*. 2014

This paper lays out a broad framework for understanding the importance of goals in shaping RL in humans, and proposes a complex network of cortical and sub-cortical regions supporting goal-based RL, explicating this framework in the context of an associative rather than TD-based picture of biological RL.

7. O'Doherty JP: **Reward representations and reward-related learning in the human brain: insights from neuroimaging**. *Curr Opin Neurobiol* 2004, **14**:769-776.

8. Solway A, Botvinick MM: **Goal-directed decision making as probabilistic inference: a computational framework and potential neural correlates**. *Psychol Rev* 2012, **119**:120-154.

9. Keller GB, Mrsic-Flogel TD: **Predictive processing: a canonical cortical computation**. *Neuron* 2018, **100**:424-435.

10. Doya K: **What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?** *Neural Netw* 1999, **12**:961-974.

11. Klaus A, Martins GJ, Paixao VB, Zhou P, Paninski L, Costa RM: **The Spatiotemporal organization of the striatum encodes action space**. *Neuron* 2017, **95**:1171-1180.e7.

12. Markowitz JE, Gillis WF, Beron CC, Neufeld SQ, Robertson K, Bhagat ND, Peterson RE, Peterson E, Hyun M, Linderman SW *et al.*: **The striatum organizes 3D behavior via moment-to-moment action selection**. *Cell* 2018, **174**:44-58.e17.

13. Balleine BW, O'Doherty JP: **Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action**. *Neuropsychopharmacology* 2010, **35**:48-69.

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

15. O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ: **Dissociable roles of ventral and dorsal striatum in instrumental conditioning**. *Science* 2004, **304**:452-454.

16. Chen C, Omiya Y, Yang S: **Dissociating contributions of ventral and dorsal striatum to reward learning**. *J Neurophysiol* 2015, **114**:1364-1366.

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

18. • Dabney W, Kurth-Nelson Z, Uchida N, Starkweather CK, Hassabis D, Munos R, Botvinick M: **A distributional code for value in dopamine-based reinforcement learning**. *Nature* 2020, **577**:671-675

This recent study revisited canonical reward prediction error theory of dopamine and propose that instead of a single scalar quantity, brain represents future rewards as a probability distribution, multiplexing possible outcomes.

19. Grillner S, Robertson B, Stephenson-Jones M: **The evolutionary origin of the vertebrate basal ganglia and its role in action selection**. *J Physiol* 2013, **591**:5425-5431.

20. Gadagkar V, Puzerey PA, Chen R, Baird-Daniel E, Farhang AR, Goldberg JH: **Dopamine neurons encode performance error in singing birds**. *Science (80-)* 2016, **354**:1278-1282.

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

22. Schultz W: **Multiple reward signals in the brain**. *Nat Rev Neurosci* 2000, **1**:199-207.

23. Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ: **Canonical microcircuits for predictive coding**. *Neuron* 2012, **76**:695-711.

24. Rikhye RV, Wimmer RD, Halassa MM: **Toward an integrative theory of thalamic function**. *Annu Rev Neurosci* 2018, **41**:163-183.

25. FitzGerald THB, Friston KJ, Dolan RJ: **Characterising reward outcome signals in sensory cortex**. *Neuroimage* 2013, **83**:329-334.

26. Shuler MG, Bear MF: **Reward timing in the primary visual cortex**. *Science (80-)* 2006, **311**:1606-1609.

27. Stanisor L, van der Togt C, Pennartz CMA, Roelfsema PR: **A unified selection signal for attention and reward in primary visual cortex**. *Proc Natl Acad Sci U S A* 2013, **110**:9136-9141.

28. •• Ramesh RN, Burgess CR, Sugden AU, Gyetvan M, Andermann ML: **Intermingled ensembles in visual association cortex encode stimulus identity or predicted outcome**. *Neuron* 2018, **100**:900-915.e9

This study applied longitudinal Ca^{2+} imaging of cortical neurons in the lateral visual association area across weeks before and after context switching and observed correlated ensembles of neurons that tracked predictive value of a stimulus.

29. McNiel DB, Choi JS, Hessburg JP, Francis JT: **Reward value is encoded in primary somatosensory cortex and can be decoded from neural activity during performance of a psychophysical task**. *Conf Proc. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Conf 2016* 2016:3064-3067.

30. •• Banerjee A, Parente G, Teutsch J, Lewis C, Voigt FF, Helmchen F: **Value-guided remapping of sensory cortex by lateral orbitofrontal cortex**. *Nature* 2020, **585**:245-250

This recent study investigated the interaction between primary sensory cortex and OFC in mice during a tactile reversal learning task. They find a causal role of OFC neurons encoding a value-prediction error signal during a rule-switch that is further conveyed to the sensory cortex for sensory remapping and flexible updating of stimulus-outcome associations.

31. Brosch M, Selezneva E, Scheich H: **Representation of reward feedback in primate auditory cortex**. *Front Syst Neurosci* 2011, **5**:5.

32. Hira R, Ohkubo F, Masamizu Y, Ohkura M, Nakai J, Okada T, Matsuzaki M: **Reward-timing-dependent bidirectional modulation of cortical microcircuits during optical single-neuron operant conditioning**. *Nat Commun* 2014, **5**:5551.

33. Grossberg S: **Desirability, availability, credit assignment, category learning, and attention: cognitive-emotional and working memory dynamics of orbitofrontal, ventrolateral, and dorsolateral prefrontal cortices**. *Brain Neurosci Adv* 2018, **2** 239821281877217.

34. Le Merre P, Esmaeili V, Charrière E, Galan K, Salin P-A, Petersen CCH, Crochet S: **Reward-based learning drives rapid sensory signals in medial prefrontal cortex and dorsal hippocampus necessary for goal-directed behavior**. *Neuron* 2018, **97**:83-91.e5.

35. •• Rikhye RV, Gilra A, Halassa MM: **Thalamic regulation of switching between cortical representations enables cognitive flexibility**. *Nat Neurosci* 2018, **21**:1753-1763

This study investigated the interaction between PFC and MD thalamus in mice and finds a causal role of distinct MD neuronal populations in flexible prefrontal representation of task-related activity upon context-switching.

36. Malagon-Vina H, Ciocchi S, Passecker J, Dorffner G, Klausberger T: **Fluid network dynamics in the prefrontal cortex during multiple strategy switching**. *Nat Commun* 2018, **9**:309.

37. •• Ferenczi EA, Zalocusky KA, Liston C, Grosenick L, Warden MR, Amatya D, Katovich K, Mehta H, Patenaude B, Ramakrishnan C *et al.*: **Prefrontal cortical regulation of brainwide circuit dynamics and reward-related behavior**. *Science* 2016, **351** aac9698

This study combined optogenetic stimulation with whole-brain functional magnetic resonance imaging in rodents to reveal how prefrontal cortex can modulate the activity of the striatum in reinforcement learning by regulating its response to dopamine.

38. Redish AD: *The Mind within the Brain: How We Make Decisions and How Those Decisions Go Wrong*. Oxford University Press; 2013.

39. Hamilton DA, Brigman JL: **Behavioral flexibility in rats and mice: contributions of distinct frontocortical regions**. *Genes Brain Behav* 2015, **14**:4-21.

40. Rolls ET: **The orbitofrontal cortex and reward**. *Cereb Cortex* 2000, **10**:284-294.

41. Murray EA, Rudebeck PH: **Specializations for reward-guided decision-making in the primate ventral prefrontal cortex**. *Nat Rev Neurosci* 2018, **19**:404-417.

42. Howard JD, Kahnt T: **Identity prediction errors in the human midbrain update reward-identity expectations in the orbitofrontal cortex**. *Nat Commun* 2018, **9**:1611.

43. Curtis CE, D'Esposito M: **Persistent activity in the prefrontal cortex during working memory**. *Trends Cogn Sci* 2003, **7**:415-423.

44. Stalnaker TA, Cooch NK, McDannald MA, Liu T-L, Wied H, Schoenbaum G: **Orbitofrontal neurons infer the value and identity of predicted outcomes**. *Nat Commun* 2014, **5**:3926.

45. Izquierdo A, Darling C, Manos N, Pozos H, Kim C, Ostrander S, Cazares V, Stepp H, Rudebeck PH: **Basolateral amygdala lesions facilitate reward choices after negative feedback in rats**. *J Neurosci* 2013, **33**:4105-4109.

46. Baltz ET, Yalcinbas EA, Renteria R, Gremel CM: **Orbital frontal cortex updates state-induced value change for decision-making**. *eLife* 2018, **7**.

47. Gremel CM, Costa RM: **Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions**. *Nat Commun* 2013, **4**.

48. Ward RD, Winiger V, Kandel ER, Balsam PD, Simpson EH: **Orbitofrontal cortex mediates the differential impact of signaled-reward probability on discrimination accuracy**. *Front Neurosci* 2015, **9**:230.

49. Wang PY, Boboila C, Chin M, Higashi-Howard A, Shamash P, Wu Z, Stein NP, Abbott LF, Axel R: **Transient and persistent representations of odor value in prefrontal cortex**. *Neuron* 2020, **108**.

50. •• Zhou J, Jia C, Montesinos-Cartagena M, Gardner MPH, Zong W, Schoenbaum G: **Evolving schema representations in orbitofrontal ensembles during learning**. *Nature* 2020 http://dx.doi.org/10.1038/s41586-020-03061-2

This recent study utilised single-unit recordings and investigated how learning of an odor sequence problem is accompanied with the formation and evolution of a neural schema in the OFC.

51. Ballesta S, Shi W, Conen KE, Padoa-Schioppa C: **Values encoded in orbitofrontal cortex are causally related to economic choices**. *Nature* 2020, **588**:450-453.

52. • Knudsen EB, Wallis JD: **Closed-loop theta stimulation in the orbitofrontal cortex prevents reward-based learning**. *Neuron* 2020, **106**:537-547.e4

Using closed-loop control, this study shows theta oscillation in OFC are critically important for reward-guided learning and they are driven by hippocampal theta-frequency oscillation.

53. Groman SM, Keistler C, Keip AJ, Hammarlund E, DiLeone RJ, Pittenger C, Lee D, Taylor JR: **Orbitofrontal circuits control multiple reinforcement-learning processes**. *Neuron* 2019, **103**:734-746.e3.

54. • Allen WE, Kauvar IV, Chen MZ, Richman EB, Yang SJ, Chan K, Gradinaru V, Deverman BE, Luo L, Deisseroth K: **Global representations of goal-directed behavior in distinct cell types of mouse neocortex**. *Neuron* 2017, **94**:891-907.e6

This study used large-scale two-photon and cortex-wide wide-field Ca^{2+} imaging to identify a global cortical representation of task-related information to guide behavior. Focal optogenetic inhibition of pre-motor cortex revealed a crucial role of this area in triggering this cortex-wide phenomenon.

55. • Chao ZC, Takaura K, Wang L, Fujii N, Dehaene S: **Large-scale cortical networks for hierarchical prediction and prediction error in the primate brain**. *Neuron* 2018, **0**

This paper utilized high-density electrocorticography recordings in monkeys performing an auditory 'local-global' paradigm and identified lower- and higher-level prediction-error signals in auditory cortex and anterior temporal cortex, respectively, and a prediction-update signal sent from PFC back to temporal cortex involving alpha/beta oscillations.

56. Averbeck BB, Costa VD: **Motivational neural circuits underlying reinforcement learning**. *Nat Neurosci* 2017, **20**:505-512.

57. Miller KJ, Botvinick MM, Brody CD: **Value representations in orbitofrontal cortex drive learning, not choice**. *bioRxiv* 2018 http://dx.doi.org/10.1101/245720.

58. Liu D, Deng J, Zhang Z, Zhang ZY, Sun YG, Yang T, Yao H: **Orbitofrontal control of visual cortex gain promotes visual associative learning**. *Nat Commun* 2020, **11**:1-14.

59. Schaul T, Horgan D, Gregor K, Silver D: **Universal value function approximators**. *Proc 32nd Int Conf Int Conf Mach Learn - Vol 37* 2015.

60. Emiliani V, Cohen AE, Deisseroth K, Hausser M: **All-optical interrogation of neural circuits**. *J Neurosci* 2015, **35**:13917-13926.

61. Botvinick M, Weinstein A: **Model-based hierarchical reinforcement learning and human action control**. *Philos Trans R Soc Lond B Biol Sci* 2014, **369**.

62. Ribas-Fernandes JJF, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM: **A neural signature of hierarchical reinforcement learning**. *Neuron* 2011, **71**:370-379.

63. Geddes CE, Li H, Jin X: **Optogenetic editing reveals the hierarchical organization of learned action sequences**. *Cell* 2018, **174**:32-43.e15.

64. Jin X, Costa RM: **Shaping action sequences in basal ganglia circuits**. *Curr Opin Neurobiol* 2015, **33**:188-196.

65. Herd SA, Krueger KA, Kriete TE, Huang T-R, Hazy TE, O'Reilly RC: **Strategic cognitive sequencing: a computational cognitive neuroscience approach**. *Comput Intell Neurosci* 2013, **2013** 149329.

66. Stachenfeld KL, Botvinick MM, Gershman SJ: **The hippocampus as a predictive map**. *bioRxiv* 2017 http://dx.doi.org/10.1101/097170.

67. Gershman SJ: **The successor representation: its computational logic and neural substrates**. *J Neurosci* 2018, **38**:7193-7200.

68. Wilson RC, Takahashi YK, Schoenbaum G, Niv Y: **Orbitofrontal cortex as a cognitive map of task space**. *Neuron* 2014, **81**:267-279.

69. Schuck NW, Cai MB, Wilson RC, Niv Y: **Human orbitofrontal cortex represents a cognitive map of state space**. *Neuron* 2016, **91**:1402-1412.

70. Rikhye RV, Gothoskar N, Guntupalli JS, Dedieu A, Lázaro-Gredilla M, George D: **Learning cognitive maps for vicarious evaluation**. *bioRxiv* 2019 http://dx.doi.org/10.1101/864421.

71. •• Wang JX, Kurth-Nelson Z, Kumaran D, Tirumala D, Soyer H, Leibo JZ, Hassabis D, Botvinick M: **Prefrontal cortex as a meta-reinforcement learning system**. *Nat Neurosci* 2018, **21**:860-868

This paper trains a recurrent neural network using reinforcement learning on a wide range of tasks, demonstrating that the network can 'learn to learn' by learning to implement a fast reinforcement learning system in the activity dynamics of the recurrent network, and argues that a similar mechanism could allow the dopamine based reinforcement learning system in the striatum to train the prefrontal cortex to implement its own reinforcement learning algorithm.

72. Kriegeskorte N, Douglas PK: **Cognitive computational neuroscience**. *Nat Neurosci* 2018, **21**:1148-1160.

73. Hassabis D, Kumaran D, Summerfield C, Botvinick M: **Neuroscience-inspired artificial intelligence**. *Neuron* 2017, **95**:245-258.

74. •• Sweis BM, Abram SV, Schmidt BJ, Seeland KD, MacDonald AW, Thomas MJ, Redish AD: **Sensitivity to "sunk costs" in mice, rats, and humans**. *Science (80-)* 2018, **361**:178-181

This paper uses a cross-species foraging task in mice, rats and humans to investigate sunk-cost bias and reports similar sensitivities to temporal sunk costs in decision-making across the species. Sensitivity to time invested accrued only after an initial decision was made, and the amount of time spent waiting increases commitment to continuing reward pursuit in these tasks. Such cross-species behavioral features may be interesting computationally as signatures of deviations from classical model free reinforcement learning and may be accessible to detailed physiological investigation in animals.