

Reward-Based Learning, Model-Based and Model-Free

Quentin J. M. Huys^{a,b*}, Anthony Cruickshank^c and Peggy Seriès^c

^aTranslational Neuromodeling Unit, Institute of Biomedical Engineering, ETH Zürich and University of Zürich, Zürich, Switzerland

^bDepartment of Psychiatry, Psychotherapy and Psychosomatics, Hospital of Psychiatry, University of Zürich, Zürich, Switzerland

^cInstitute of Adaptive and Neural Computation, University of Edinburgh, Edinburgh, UK

Definition

Reinforcement learning (RL) techniques are a set of solutions for optimal long-term action choice such that actions take into account both immediate and delayed consequences. They fall into two broad classes. Model-based approaches assume an explicit model of the environment and the agent. The model describes the consequences of actions and the associated returns. From this, optimal policies can be inferred. Psychologically, model-based descriptions apply to goal-directed decisions, in which choices reflect current preferences over outcomes. Model-free approaches forgo any explicit knowledge of the dynamics of the environment or the consequences of actions and evaluate how good actions are through trial-and-error learning. Model-free values underlie habitual and Pavlovian conditioned responses that are emitted reflexively when faced with certain stimuli. While model-based techniques have substantial computational demands, model-free techniques require extensive experience.

Detailed Description

Theory

Reinforcement Learning

Formally, reinforcement learning (RL; Sutton and Barto 1998) describes a type of solution to Markov decision process (MDP) problems which are defined by a tuple $\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}$, and π :

- \mathcal{S} : a set of states $s \in \mathcal{S}$
- \mathcal{A} : a set of actions $a \in \mathcal{A}$
- $\mathcal{T}(s'|s, a)$: the transition function maps each state-action pair to a distribution over successor states s' , with $s, s' \in \mathcal{S}$; $a \in \mathcal{A}$, and $\sum_{s'} \mathcal{T}(s'|s, a) = 1$
- $\mathcal{R}(s, a, s') \rightarrow r$: the reinforcement function mapping state-action-successor state triples to a scalar return r

The goal is to determine a policy $a \leftarrow \pi(s)$ that maps each state to the action maximizing the total expected future return of actions a in state s

*Email: qhuys@cantab.net

$$a^* \leftarrow \operatorname{argmax}_a Q(s, a) \quad \text{where} \quad Q(s, a) = \mathbb{E} \left[\sum_{t=0}^{\infty} r_t \mid s, a \right] \quad (1)$$

where the sum over the future returns results from the fact that choices lead both to immediate returns but also have longer-term consequences.

The sum in Eq. 1 may not be finite. For this reason, it is often replaced by the discounted total expected reward $\mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s, a \right]$ with the discount factor $0 \leq \gamma \leq 1$. The discount factor sets the relative importance of immediate and future rewards: $\gamma = 0$ means that only the next reward is considered whereas $\gamma = 1$ considers all rewards to have equal importance no matter how far in the future they occur.

Model-Based RL

Model-based RL assumes knowledge of the transition matrix \mathcal{T} , the reward function \mathcal{R} , and the state and action spaces \mathcal{S} and \mathcal{A} which define the model of the world. This means that the expectation in Eq. 1 can be written explicitly in terms of \mathcal{T} and \mathcal{R} as the *Bellman equation* (Bellman 1957):

$$Q(s, a) = \sum_{s'} \mathcal{T}(s' | a, s) [\mathcal{R}(s, a, s') + V(s')] \quad (2)$$

$$\text{with} \quad V(s') = \max_{a'} Q(s', a'). \quad (3)$$

The Q value is the long-run expected return for taking action a in state s . The optimal policy maps each state to the action with the highest Q value:

$$\pi^*(s) \leftarrow \operatorname{argmax}_a Q(s, a). \quad (4)$$

Equation 3 represents a recursive definition of a decision-tree of width w (determined by the number of actions $|\mathcal{A}|$ and the size of the state-space reached by these actions). The computational cost of simple tree search is $\mathcal{O}(w^d)$ where d is the depth of the tree (see Fig. 1 for an example). Although dynamic programming methods such as policy iteration reduce this cost to $\mathcal{O}(|\mathcal{S}|^3)$, this is still computationally prohibitive for most real-life problems and additionally difficult to implement neurally as it involves matrix inversion. Psychological and neurobiological accounts of model-based RL thus emphasize sequential evaluations of decision-trees.

Model-Free RL

Model-free RL methods apply to situations where agents do not know \mathcal{T} and \mathcal{R} where the decision-trees are too complex to evaluate. They approximate the expectations in Eq. 1 by sampling from the world. *Temporal difference reinforcement learning (TDRL)* constructs estimates of state or state-action values from these samples by bootstrapping. To achieve this, the total future reward is written as the sum of the immediate reward plus the average value of the successor state:

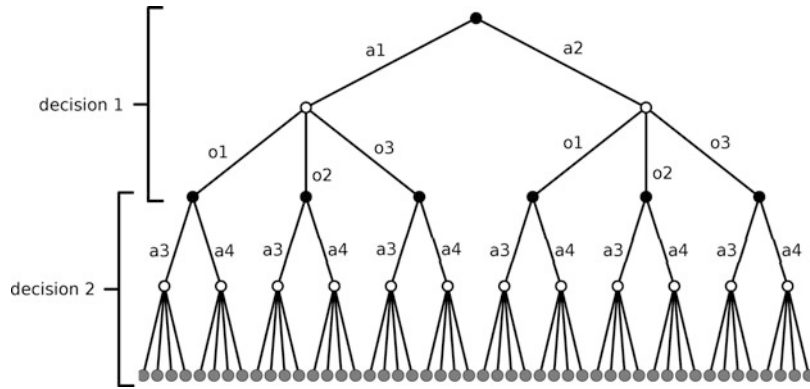


Fig. 1 Example decision-tree. The problem consists of first choosing between actions $a1$ and $a2$, each of which has three possible outcomes, followed by choosing between actions $a3$ and $a4$, each of which has another three possible outcomes that might depend on which action/outcome preceded them. Actions are shown in solid black and outcomes as empty circles. Optimal action choice requires evaluation of all the branches. In this simple problem, with a sequence of two choices, each leading to three possible outcomes, the tree has width $w = 6$, depth $d = 2$, and $w^d = 36$ branches (Adapted from Huys 2007)

$$\mathcal{V}^*(s) = \mathbb{E} \left[\sum_{t=0}^{\infty} r_t \middle| s \right] = \mathbb{E} \left[r_0 + \sum_{t=1}^{\infty} r_t \middle| s \right] = \mathbb{E}[r_0 + \mathcal{V}^*(s') | s] = \mathbb{E}[r_0 | s] + \mathbb{E}[\mathcal{V}^*(s') | s]. \quad (5)$$

For approximate values \mathcal{V} , Eq. 5 does not hold:

$$\hat{\mathcal{V}}(s) \neq \mathbb{E}[r_0 | s] + \mathbb{E}[\hat{\mathcal{V}}(s') | s]. \quad (6)$$

Letting the difference between the two sides be $\delta_{\mathcal{V}}$, one can arrive at correct values by iterative updates

$$\hat{\mathcal{V}}_{i+1}(s) \leftarrow \hat{\mathcal{V}}_i(s) + \varepsilon \delta_{\mathcal{V}} \quad (7)$$

with $0 \leq \varepsilon \leq 1$.

TDRL combines such iterative updates with sampling. Instead of evaluating the expectations, it assumes that agents can repeatedly generate actions from their (suboptimal) policy $a_t \sim \pi(s_t)$ and on the t^{th} such interaction obtain state and reward samples from the world:

$$s_{t+1} \sim \mathcal{T}(s | s_t, a_t) \quad (8)$$

$$r_t \sim \mathcal{R}(s_t, a_t, s_{t+1}). \quad (9)$$

These samples are used to approximate the expectations, letting:

$$\delta_t = r_t + \hat{\mathcal{V}}_t(s_{t+1}) - \hat{\mathcal{V}}_t(s_t) \quad (10)$$

$$\hat{\mathcal{V}}_{t+1}(s_t) \leftarrow \hat{\mathcal{V}}_t(s_t) + \varepsilon \delta_t. \quad (11)$$

A similar approach can be applied to learning state-action values (Watkins and Dayan 1992). Thus, while model-based RL methods prospectively predict the consequences of actions based on an

understanding of the structure of the world, model-free methods retrospectively approximate these based on past experience. Nevertheless, under certain situations, model-free methods have strong convergence guarantees (Bertsekas and Tsitsiklis 1996; Sutton and Barto 1998; Puterman 2005). Policies π are often in turn formalized as parametric functions of the value functions \mathcal{V} or \mathcal{Q} themselves, although this may break certain guarantees (Bertsekas and Tsitsiklis 1996).

One biologically important variation of a reinforcement learning algorithm is the **actor-critic** (Barto et al. 1983). The critic uses TD to estimate the value $\mathcal{V}_t(s)$ for states while the actor maintains the policy used to select actions. After each action a_t , the critic calculates the prediction error and sends it to the actor. A positive prediction error indicates that the action improved the potential for future rewards, and the tendency to select the action should be increased. An example of using the prediction error is to select actions based on the Gibbs softmax method

$$\pi_t(s, a) = \frac{e_t^{p(s,a)}}{\sum_{a'} e_t^{p(s,a')}} \quad (12)$$

where $p_t(s, a)$ defines the “propensity” to take action a in state s . These propensities are updated by the prediction error $p_t(s, a) \leftarrow p_{t-1}(s, a) + \varepsilon \delta_t$.

Sampling and Computational Costs

The algorithms discussed so far suffer either from catastrophic computational requirements or from equally drastic dependence on extensive sampling in realistic environments. Solutions to these drawbacks fall into four general categories: (1) subdivision into smaller subtasks (possibly each having their own subgoal; c.f. Dietterich 1999; Sutton et al. 1999), (2) pruning of the decision-tree (c.f. Knuth and Moore 1975; Huys et al. 2012), and (3) approximations (e.g., neural networks for function approximation (Sutton and Barto 1998) or structured representations (Boutilier et al. 1995)) and sampling techniques (Kearns and Singh 2002; Kocsis and Szepesvári 2006). If the states \mathcal{S} are not fully observable, the problem becomes a partially observable MDP (Kaelbling et al. 1998), which presents substantial additional complexities.

The strengths of model-based and model-free computations can also be combined to offset their mutual weaknesses. In *Dyna-Q* (Sutton 1990), samples as in Eq. 9 are generated from the agent’s internal estimates of \mathcal{T} and \mathcal{R} to update model-free values. Conversely, model-free state values can be substituted for subtrees to reduce the size of decision-trees (e.g., Campbell et al. 2002).

Behavior

Model-based and model-free accounts of behavior were held to be incompatible for much of the last century (Hull 1943; Tolman 1948). Tolman (1948) classically showed that animals were able to acquire and use internal maps of environments. They were able to use these internal models and did not necessarily require pairings of actions and outcomes (Fig. 2). Indeed, key signatures of both systems can be discerned within individual animals’ (Balleine and Dickinson 1994; Killcross and Coutureau 2003; Yin et al. 2004, 2005) and humans’ (Valentin et al. 2007; Daw et al. 2011) behavior and neurobiology. These signatures reflect central differences in their utilization of information. For a discussion, see Daw et al. (2005), Dayan and Berridge (2013), and Huys et al. (In Press).

In instrumental paradigms, particular actions a are reinforced in the presence of certain stimuli or situations s . These experiments are modeled using $\mathcal{Q}(a, s)$ values. In Pavlovian paradigms, stimuli s lead to reinforcements independent of subjects’ actions. These paradigms are modeled using stimulus values $\mathcal{V}(s)$. Importantly, there can be model-based and model-free versions of both, leading

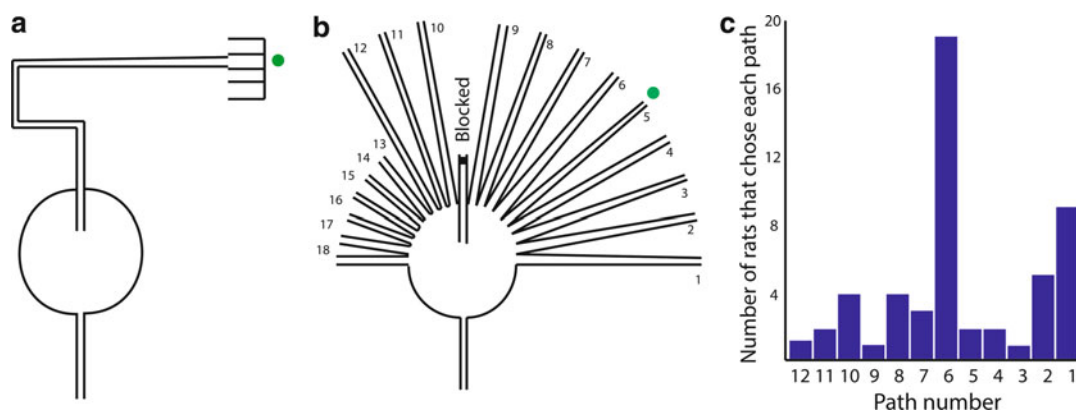


Fig. 2 Early experiment used to argue that rats can build and use spatial representations (Figure after Tolman 1948). (a) Rats were first trained to find a food source located below the *green point* (a light). (b) After training, the rats were placed in the same starting position at the bottom of the maze, but found their usual route blocked. Instead, they now had multiple alternative arms they could run down. (c) Histogram of arm the rats chose to run down first

to a quartet of values $\mathcal{V}^{\text{MF}}(s)$, $\mathcal{V}^{\text{MB}}(s)$, $Q^{\text{MF}}(s, a)$, and $Q^{\text{MB}}(s, a)$. Both model-free values $\mathcal{V}^{\text{MF}}(s)$ and $Q^{\text{MF}}(s, a)$ are *scalar* representations that change *slowly*. These two features account for its key behavioral signatures.

The consequences of the *scalar* nature of model-free values is most clearly seen in Pavlovian scenarios, where $\mathcal{V}^{\text{MF}}(s)$ reflect only the magnitude of reinforcements but not other aspects such as whether an action was rewarded by food or water. One paradigmatic example are blocking experiments (Kamin 1969). In these, learning the reward association of a stimulus “B” in a compound “AB” is prevented if “A” already fully predicts the reward. When the reward is fully predicted, no prediction error occurs. Hence, model-free values are not updated and hence no learning occurs.

Thus, if the model-free system makes no prediction about certain aspects of stimuli, then shifts in these aspects should not lead to learning. In transreinforcer blocking, animals treat a reward reduction and delivery of a shock punishment as equivalent (Dickinson and Dearing 1979), arguing for a linear and unitary representation of rewards and punishments as encapsulated in the single value r in Eq. 11. In Pavlovian unblocking, animals similarly can show an insensitivity towards shifts between rewards of equal magnitude but different modality (e.g., water and food; McDannald et al. 2011), showing that only the reward value, but not its other sensory features, is encoded. As a scalar value, model-free values can, however, replace reinforcements and be approached (if positive; Dayan et al. 2006) or avoided (if negative; Guitart-Masip et al. 2011). In conditioned reinforcement experiments, behavior is motivated by stimuli associated with the rewards (i.e., having positive model-free value $\mathcal{V}^{\text{MF}}(s)$) even in the absence of the rewards themselves (Bouton 2006). This can be captured by actor-critic models (Barto et al. 1983). By the same argument, model-free state or stimulus values $\mathcal{V}^{\text{MF}}(s)$ can also influence the vigor with which ongoing actions are performed (Pavlovian-instrumental transfer; Huys et al. 2011). These three features are also central to the notion of incentive value (McClure et al. 2003).

Model-free values change slowly over time as they rely on iterative updating (Eqs. 7 and 11). The consequences of this have been mainly examined in instrumental settings (though see Schoenbaum et al. (2009) and Robinson and Berridge (2013) for Pavlovian examples). The paradigmatic example is outcome devaluation (Fig. 3). On the very first trial after the devaluation, the model-free system would have had no opportunity to update the $Q^{\text{MF}}(s, a)$ values via prediction errors δ and hence would predict continued responding. Conversely, by considering the now undesired outcome of

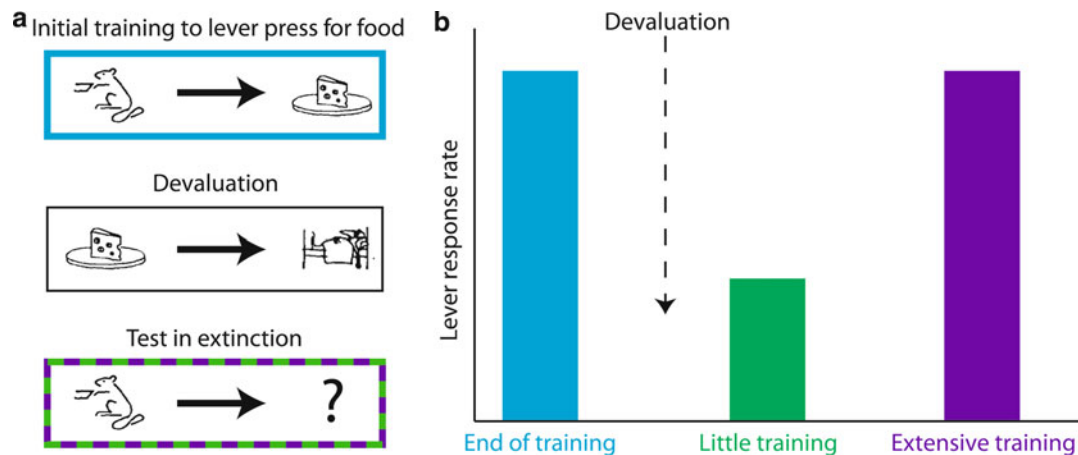


Fig. 3 Devaluation experiments. **(a)** Animals are first reinforced to press a lever for a particular food for either a brief period of time or for a long period. This food is then devalued, for instance by satiation or by pairing with illness. Animals are then given the opportunity to press the lever again, though in the absence of any food outcomes (in extinction). **(b)** After brief initial training, animals will refuse to press the lever (*green bar*), but after extensive training, they will press the lever (*purple bar*) at the same rate as at the end of training (*blue bar*) despite refusing to consume the food if given the opportunity (Figure adapted from Balleine and Dickinson 1994)

actions, model-based evaluation should lead to a reduction in lever pressing on the very first trial after the devaluation. Accounts of the shift from early model-based and goal-directed to later model-free and habitual behavior rely on their statistical properties (Daw et al. 2005) or the trade-off between the cost of cognition and the value of improved choices (Keramati et al. 2011).

Neurobiology

The component of model-free learning best understood is the representation of the temporal prediction error δ . Interpreting earlier work by Schultz and Romo (1990), Montague et al. (1996) pointed out that the phasic firing of dopaminergic midbrain neurons corresponds closely to the positive portion of the prediction error δ . This has been extensively validated with single-electrode recordings (even in humans; Zaghoul et al. 2009), functional neuroimaging (D'Ardenne et al. 2008), cyclic voltammetry (Day et al. 2007), with optogenetic manipulations (Steinberg et al. 2013), and in diseases of the dopamine neurons (Frank et al. 2004). This is true both in Pavlovian (Waelti et al. 2001; Flagel et al. 2011) and instrumental scenarios (Morris et al. 2006; Roesch et al. 2007). These phasic prediction errors are not just a linear reflection of the magnitude and probability of the expected reward (Tobler et al. 2005; Bayer and Glimcher 2005) but also of the summed long-term future rewards (Schultz et al. 1997; Enomoto et al. 2011). Dopamine neurons have a low firing baseline and therefore appear to represent the negative portion of the prediction errors δ by the length of the pause in firing (Bayer et al. 2007). Phasic firing covaries with the development of behavioral responses (Waelti et al. 2001; Flagel et al. 2011) and can causally drive learning (Steinberg et al. 2013). Furthermore, pharmacological manipulations of dopamine alter the behavioral expression of model-free versus model-based behaviors (Nelson and Killcross 2006; Wunderlich et al. 2012).

In comparison, the neural location where prediction errors are summated into model-free values is much less well understood, although multiple parts of the affective neural circuitry appear to be involved, from the ventral (Cardinal et al. 2002; Corbit and Balleine 2005b; McDannald et al. 2011) and dorsal portions of the striatum (Yin et al. 2004, 2005) and the ventromedial prefrontal cortex

(Killcross and Coutureau 2003; Smith and Graybiel 2013) to the amygdala (Corbit and Balleine 2005a).

Similarly, the neural bases of the model-based system are also poorly understood. Depending on the nature of the structure represented in \mathcal{T} , different neural substrates will be required. Hence, there is a priori no reason to expect a unitary representation of a single model-based system. However, particular features of the system can probably be pinpointed. For instance, learning about a stimulus-stimulus transition matrix recruits the posterior parietal cortex (Gläscher et al. 2010), while model-based expectations of stimulus value involve the ventromedial prefrontal cortex (Hampton et al. 2006; Schoenbaum et al. 2009). Recordings from spatial navigation tasks in the rodent hippocampus are so far unique in yielding direct neural evidence of the implementation of sequential tree search (Johnson and Redish 2007; Pfeiffer and Foster 2013).

Psychopathology

Given the representation of a key model-free component by dopaminergic neurons, pathological excesses of dopamine have been suggested to involve a shift from model-based to model-free decision-making (Redish et al. 2008; Robbins et al. 2012; Huys et al. *In Press*). This has been clearly demonstrated in laboratory animals (Dickinson et al. 2000; Nelson and Killcross 2006), and data in humans are forthcoming (Sebold et al. *In press*). Similar arguments have been made about other disorders with a striatal component, particularly obsessive-compulsive disorders (Gillan et al. 2011, 2014), and models incorporating additional neurobiological details about the striatum can account for some of the choice patterns seen in Parkinson's disease, ADHD, and Tourette's syndrome (Maia and Frank 2011).

References

- Balleine B, Dickinson A (1994) Role of cholecystokinin in the motivational control of instrumental action in rats. *Behav Neurosci* 108(3):590–605
- Barto A, Sutton R, Anderson C (1983) Neuronlike elements that can solve difficult learning control problems. *IEEE Trans Syst Man Cybern* 13(5):834–846
- Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47(1):129–141
- Bayer HM, Lau B, Glimcher PW (2007) Statistics of midbrain dopamine neuron spike trains in the awake primate. *J Neurophysiol* 98(3):1428–1439
- Bellman RE (1957) *Dynamic programming*. Princeton University Press, Princeton
- Bertsekas DP, Tsitsiklis JN (1996) *Neuro-dynamic programming*. Athena Scientific, Belmont
- Boutillier C, Dearden R, Goldszmidt M (1995) Exploiting structure in policy construction. In: *Proceedings of the IJCAI Montreal, Quebec, Canada August 20–25, 1995*, vol 14, pp 1104–1113
- Bouton ME (2006) *Learning and behavior: a contemporary synthesis*. Sinauer, Sunderland
- Campbell M, Hoane A et al (2002) Deep blue. *Artif Intell* 134(1–2):57–83
- Cardinal RN, Parkinson JA, Lachenal G, Halkerston KM, Rudarakanchana N, Hall J, Morrison CH, Howes SR, Robbins TW, Everitt BJ (2002) Effects of selective excitotoxic lesions of the nucleus accumbens core, anterior cingulate cortex, and central nucleus of the amygdala on autoshaping performance in rats. *Behav Neurosci* 116(4):553–567
- Corbit LH, Balleine BW (2005a) Double dissociation of basolateral and central amygdala lesions on the general and outcome-specific forms of Pavlovian-instrumental transfer. *J Neurosci* 25(4):962–970

- Balleine BW, Corbit LH (2005b) Double dissociation of nucleus accumbens core and shell on the general and outcome-specific forms of Pavlovian-instrumental transfer. Program No. 71.16. 2005 Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience, 2005. Online
- D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) Bold responses reflecting dopaminergic signals in the human ventral tegmental area. *Science* 319(5867):1264–1267
- Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat Neurosci* 8(12):1704–1711
- Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* 69(6):1204–1215
- Day JJ, Roitman MF, Wightman RM, Carelli RM (2007) Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nat Neurosci* 10(8):1020–1028
- Dayan P, Berridge KC (2013) Pavlovian values. *Cogn Affect Behav Neurosci*. 2014 Mar 20. [Epub ahead of print] doi: 10.3758/s13415-014-0277-8
- Dayan P, Niv Y, Seymour B, Daw ND (2006) The misbehavior of value and the discipline of the will. *Neural Netw* 19(8):1153–1160
- Dickinson A, Dearing MF (1979) Appetitive-aversive interactions and inhibitory processes. In: Dickinson A, Boakes RA (eds) *Mechanisms of learning and motivation*. Erlbaum, Hillsdale, pp 203–231
- Dickinson A, Smith J, Mirenowicz J (2000) Dissociation of Pavlovian and instrumental incentive learning under dopamine antagonists. *Behav Neurosci* 114(3):468–483
- Dietterich TG (1999) Hierarchical reinforcement learning with the maxq value function decomposition. CoRR, cs.LG/9905014
- Enomoto K, Matsumoto N, Nakai S, Satoh T, Sato TK, Ueda Y, Inokawa H, Haruno M, Kimura M (2011) Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proc Natl Acad Sci U S A* 108(37):15462–15467
- Flagel SB, Clark JJ, Robinson TE, Mayo L, Czuj A, Willuhn I, Akers CA, Clinton SM, Phillips PEM, Akil H (2011) A selective role for dopamine in stimulus-reward learning. *Nature* 469(7328):53–57
- Frank MJ, Seeberger LC, O'Reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in Parkinsonism. *Science* 306(5703):1940–1943
- Gillan CM, Pappmeyer M, Morein-Zamir S, Sahakian BJ, Fineberg NA, Robbins TW, de Wit S (2011) Disruption in the balance between goal-directed behavior and habit learning in obsessive-compulsive disorder. *Am J Psychiatry* 168(7):718–726
- Gillan CM, Morein-Zamir S, Urcelay GP, Sule A, Voon V, Apergis-Schoute AM, Fineberg NA, Sahakian BJ, Robbins TW (2014) Enhanced avoidance habits in obsessive-compulsive disorder. *Biol Psychiatry* 75:631–638
- Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66(4):585–595
- Guitart-Masip M, Fuentemilla L, Bach DR, Huys QJM, Dayan P, Dolan RJ, Duzel E (2011) Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *J Neurosci* 31(21):7867–7875
- Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial pre-frontal cortex in abstract state-based inference during decision making in humans. *J Neurosci* 26(32):8360–8367, 6
- Hull C (1943) *Principles of behavior*. Appleton, New York
- Huys QJM (2007) *Reinforcers and control. Towards a computational etiology of depression*. PhD thesis, Gatsby Computational Neuroscience Unit, UCL, University of London

- Huys QJM, Tobler PT, Hasler G, Flagel S. The role of learning-related dopamine signals in addiction vulnerability. *Prog Neurobiol* (In Press)
- Huys QJM, Cools R, Gölzer M, Friedel E, Heinz A, Dolan RJ, Dayan P (2011) Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol* 7(4):e1002028
- Huys QJM, Eshel N, O’Nions E, Sheridan L, Dayan P, Roiser JP (2012) Bonsai trees in your head: how the Pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol* 8(3):e1002410
- Johnson A, Redish AD (2007) Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *J Neurosci* 27(45):12176–12189
- Kaelbling LP, Littman ML, Cassandra AR (1998) Planning and acting in partially observable stochastic domains. *Artif Intell* 101(1):99–134
- Kamin LJ (1969) Predictability, surprise, attention and conditioning. In: Campbell BA, Church RM (eds) *Punishment and aversive behavior*. Appleton, New York
- Kearns M, Singh S (2002) Near-optimal reinforcement learning in polynomial time. *Mach Learn* 49(2–3):209–232
- Keramati M, Dezfouli A, Piray P (2011) Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol* 7(5):e1002055
- Killcross S, Coutureau E (2003) Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb Cortex* 13(4):400–408
- Knuth D, Moore R (1975) An analysis of alpha-beta pruning. *Artif Intell* 6(4):293–326
- Kocsis L, Szepesvári C (2006) Bandit based Monte-Carlo planning. In: *Proceedings of the Machine learning: ECML 2006*, Berlin, Germany, Springer, pp 282–293
- Maia TV, Frank MJ (2011) From reinforcement learning models to psychiatric and neurological disorders. *Nat Neurosci* 14(2):154–162
- McClure SM, Daw ND, Montague PR (2003) A computational substrate for incentive salience. *Trends Neurosci* 26:423–428
- McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G (2011) Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *J Neurosci* 31(7):2700–2705
- Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J Neurosci* 16(5):1936–1947
- Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci* 9(8):1057–1063
- Nelson A, Killcross S (2006) Amphetamine exposure enhances habit formation. *J Neurosci* 26(14):3805–3812
- Pfeiffer BE, Foster DJ (2013) Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* 497(7447):74–79
- Puterman ML (2005) *Markov decision processes: discrete stochastic dynamic programming*, Wiley series in probability and statistics. Wiley-Interscience, New York
- Redish AD, Jensen S, Johnson A (2008) A unified framework for addiction: vulnerabilities in the decision process. *Behav Brain Sci* 31(4):415–437; discussion 437–487
- Robbins TW, Gillan CM, Smith DG, de Wit S, Ersche KD (2012) Neurocognitive endophenotypes of impulsivity and compulsivity: towards dimensional psychiatry. *Trends Cogn Sci* 16(1):81–91
- Robinson MJF, Berridge KC (2013) Instant transformation of learned repulsion into motivational “wanting”. *Curr Biol* 23(4):282–289

- Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nat Neurosci* 10(12):1615–1624
- Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK (2009) A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nat Rev Neurosci* 10(12):885–892
- Schultz W, Romo R (1990) Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *J Neurophysiol* 63(3):607–624
- Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275(5306):1593–1599
- Sebold M, Deserno L, Nebe S, Schad D, Garbusow M, Hägele C, Keller J, Jünger E, Kathmann N, Smolka M, Rapp MA, Schlagenhauf F, Heinz A, Huys QJM. Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology* (In Press)
- Smith KS, Graybiel AM (2013) A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron* 79(2):361–374
- Steinberg EE, Keiflin R, Boivin JR, Witten IB, Deisseroth K, Janak PH (2013) A causal link between prediction errors, dopamine neurons and learning. *Nat Neurosci* 16(7):966–973
- Sutton R (1990) Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In: Proceedings of the seventh international conference on machine learning, Austin, Texas, USA, vol 216, p 224
- Sutton RS, Barto AG (1998) Reinforcement learning: an introduction, Computation and machine learning. The MIT Press, Cambridge, MA
- Sutton RS, Precup D, Singh S et al (1999) Between mdps and semi-mdps: a framework for temporal abstraction in reinforcement learning. *Artif Intell* 112(1):181–211
- Tobler PN, Fiorillo CD, Schultz W (2005) Adaptive coding of reward value by dopamine neurons. *Science* 307(5715):1642–1645
- Tolman EC (1948) Cognitive maps in rats and men. *Psychol Rev* 55(4):189–208
- Valentin VV, Dickinson A, O’Doherty JP (2007) Determining the neural substrates of goal-directed learning in the human brain. *J Neurosci* 27(15):4019–4026
- Waelti P, Dickinson A, Schultz W (2001) Dopamine responses comply with basic assumptions of formal learning theory. *Nature* 412(6842):43–48
- Watkins C, Dayan P (1992) Q-learning. *Mach Learn* 8(3):279–292
- Wunderlich K, Smittenaar P, Dolan RJ (2012) Dopamine enhances model-based over model-free choice behavior. *Neuron* 75(3):418–424
- Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur J Neurosci* 19(1):181–189
- Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. *Eur J Neurosci* 22(2):513–523
- Zaghloul KA, Blanco JA, Weidemann CT, McGill K, Jaggi JL, Baltuch GH, Kahana MJ (2009) Human substantia nigra neurons encode unexpected financial rewards. *Science* 323(5920):1496–1499