

(Reinforcement?) Learning to Forage Optimally

Nils Kolling & Thomas Akam

Highlights:

- Decision neuroscience is increasingly studying ecologically inspired foraging tasks.
- Human behaviour in foraging tasks is not well explained by model-free reinforcement learning.
- Humans extrapolate trends in reward rate trajectories to guide foraging choices.
- Model-based average reward reinforcement learning may support patch foraging.

Abstract: Foraging effectively is critical to the survival of all animals and this imperative is thought to have profoundly shaped brain evolution. Decisions made by foraging animals often approximate optimal strategies, but the learning and decision mechanisms generating these choices remain poorly understood. Recent work with laboratory foraging tasks in humans suggest their behaviour is poorly explained by model-free reinforcement learning, with simple heuristic strategies better describing behaviour in some tasks, and in others evidence of prospective prediction of the future state of the environment. We suggest that model-based average reward reinforcement learning may provide a common framework for understanding these apparently divergent foraging strategies.

Introduction: For decades ecologists have worked to understand the sensitivity of natural foraging to the costs and benefits of different behaviours [1,2]. In parallel, experimental psychologists and neuroscientists have worked towards understanding learning and decision making and their neural substrates in the laboratory. Reinforcement learning (RL), a field of machine learning concerned with selecting actions to maximise reward, has deeply and reciprocally influenced the laboratory study of animal learning, providing a powerful framework for understanding how adaptive behaviours can be learnt by trial and error, and insight into neural reward signals [3–5]. Decision tasks inspired by foraging have increasingly attracted the attention of neuroscientists [6–14], in part because the ecological importance of foraging is thought to have profoundly shaped neural decision making systems [15,16].

However, foraging decisions present distinctive features that differ from tasks typically used in neuroscience: Foragers search their environment and on encountering a resource patch or prey item must choose between investing time in exploiting it or continuing their search. Time spent on a given encounter is time lost to searching for richer picking elsewhere, so the opportunity cost of time is a key decision variable. The returns from different prey types have their own dynamics; decreasing as a tree is depleted of berries, increasing as the richest part of a patch is located, or requiring sustained work processing a prey for an eventual return. Deciding when to move on can therefore be as important to efficient foraging as deciding what to engage.

It remains unclear to what extent foraging behaviours utilise specialised decision heuristics or general reinforcement learning mechanisms to approximate optimal strategies. Here we examine this question in the context of patch foraging, a canonical optimal foraging behaviour where recent laboratory studies have shed light on learning mechanisms.

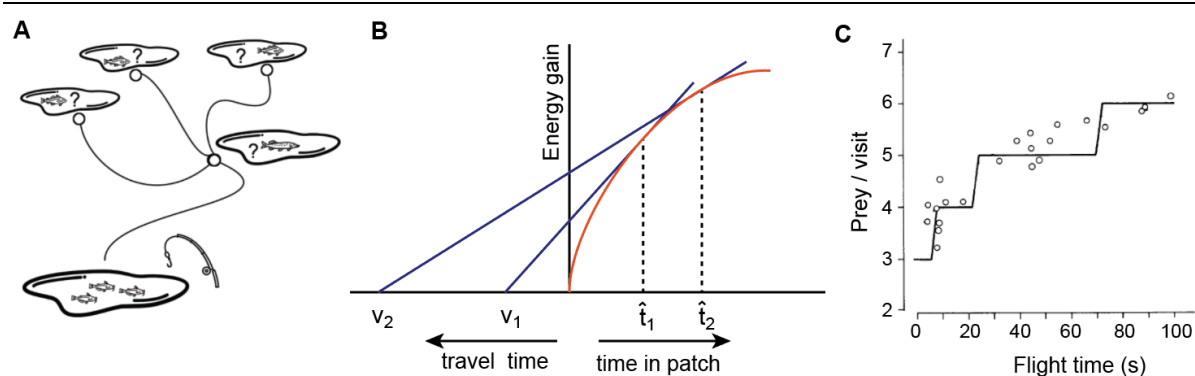


Figure 1: Patch leaving and the Marginal Value Theorem. **A)** A forager which exploits localised resource patches such as pools with fish must decide how long to stay in each patch before moving on. **B)** Diagram illustrating marginal value theory in the case where there is a single patch type for which the energy gained increases with time spent in a patch (red line) but the rate of increase slows as the patch becomes depleted. Energy gain per unit foraging time is given by the energy gain per patch divided by the travel time plus time in patch – i.e. the gradient of the blue lines. The rate of energy gain is maximised by leaving the patch when the rate of gain within the patch equals the average rate achieved in the environment, i.e. when the blue line forms a tangent to the red line. Longer travel times give longer optimal foraging time within patches, as illustrated by the two different travel times with corresponding optimal patch leaving times. **C)** Starlings were trained to collect mealworms from a feeder which delivered them at progressively longer intervals to emulate a depleting patch. Birds remained longer at the feeder, collecting more prey per visit, when the feeder was further from their nest (open circles), consistent with predictions of an MVT model (line). Adapted from [17].

Patch foraging and the Marginal Value Theory:

Optimal foraging theory assumes that foraging behaviours have been shaped through natural selection to optimise returns in the face of environmental and physiological constraints and costs [1,2]. Patch foraging models consider an animal which moves between localised resource patches, e.g. pools containing fish, harvesting resources from one patch before moving on to another (Figure 1). The longer the animal stays in a patch the more depleted the patch becomes and the slower the animal accrues food, but moving between patches costs time. How long should the animal remain in the current patch? Under the models simplifying assumptions, energy gain per unit time is maximised if the animal leaves the patch when the rate of return within it drops below the average rate of return in the environment [1,18], a result termed Marginal Value Theory (MVT). MVT has proved a powerful framework for understanding qualitative and quantitative aspects of patch leaving decisions in the field and laboratory [17,1,2,6,10].

Learning to forage

An important question regarding MVT like patch foraging is how the animal learns the properties of the current patch and broader environment [19,2,6,10,13]. A central challenge in learning from rewards is that extended sequences of actions may intersperse behaviourally salient outcomes, so credit assignment can be hard. Reinforcement learning (RL) is a framework for thinking about this problem which considers an agent that receives at each time-step information about the state of the world and a scalar valued reward signal, and must learn to select actions which maximise the long run accrued reward [4]. A diversity of RL algorithms exist, but a feature unifying many is estimation of the values of states and/or actions, typically defined as the summed discounted future reward they are expected to lead to (Box 1).

RL algorithms can be divided into two broad classes on the basis of what they learn about the world. Model-free algorithms learn value estimates directly from past experience, updating them based on reward prediction errors (Box 1). Model-based algorithms instead learn to predict the next state and immediate reward that follows each action, and use this model to work out long-run values, a process termed planning. This makes model-based RL flexible – when a change is detected in one part of the environment the consequences of this for decisions elsewhere can be evaluated using the model, whereas model-free methods must discover these consequences through experience. However forward planning is expensive in time and computational resources because the decision tree typically grows exponentially with planning depth into the future, yet shallow planning may fail to identify important long term consequences of an action.

An influential contemporary view of instrumental learning is that both model-free and model-based RL are used in parallel, instantiated in partially separate but closely interacting neural systems, to take advantage of their complementary strengths [20–22]. This was proposed on the basis of dissociations identified in rodent brain lesion studies between areas needed for goal-directed and habitual actions [23–26], and has since proved fruitful for understanding behaviour and brain activity in associative learning experiments [27–30] and multi-step decision tasks [31–36].

In spite of this general utility, It remains unclear whether the same reinforcement learning mechanisms can explain patterns of foraging, and accounts of learning in foraging behaviours have often focused on simpler learning and decision heuristics. While MVT is a description of the optimal solution to a foraging problem rather than a learning rule to achieve it, MVT does suggest a simple learning and decision rule: Estimate the environmental average reward rate by averaging the experienced reward rate over a timescale spanning encounters with many patches, compare this with the current reward rate in the patch, leave the patch when the reward rate within it drops below the environmental average [1,10,37,19]. In practice, the reward rate within a patch may fluctuate, so some averaging over recent rewards may be required to estimate the current rate. Following [1] we term this strategy the Marginal Value Rule.

The Marginal Value Rule makes good patch leaving decisions in environments where the reward rate within each patch is monotonically decreasing with time spent in the patch [19]. However, it makes poor decisions in environments where the current reward rate in a patch may be too pessimistic an estimator of the future reward rate within it, because it leaves patches prematurely compared to the optimal strategy [38,37,1]. This may be the case if patches have an initially increasing rate of return, for example as the forager locates the richest part of the patch, or if returns within the patch are stochastic so an estimate of patch quality can only be obtained over sustained sampling, potentially combined with prior learning about the statistics of patch types in the environment [38,37,39,1]. In such cases the expected future reward rate in the patch, rather than the current reward rate, is central to the optimal patch leaving strategy [38,37,1] (Box 2).

The Marginal Value Rule differs substantially from typical RL algorithms in that while RL algorithms typically learn and compare action values to reach a decision, i.e. the long run rewards each action is expected to lead to (Box 1), the Marginal Value Rule compares the current reward rate with the environmental average reward rate. This raises the question of whether learning in foraging behaviours is better describes by the Marginal Value Rule or general purpose RL algorithms?

Human foraging in the Lab

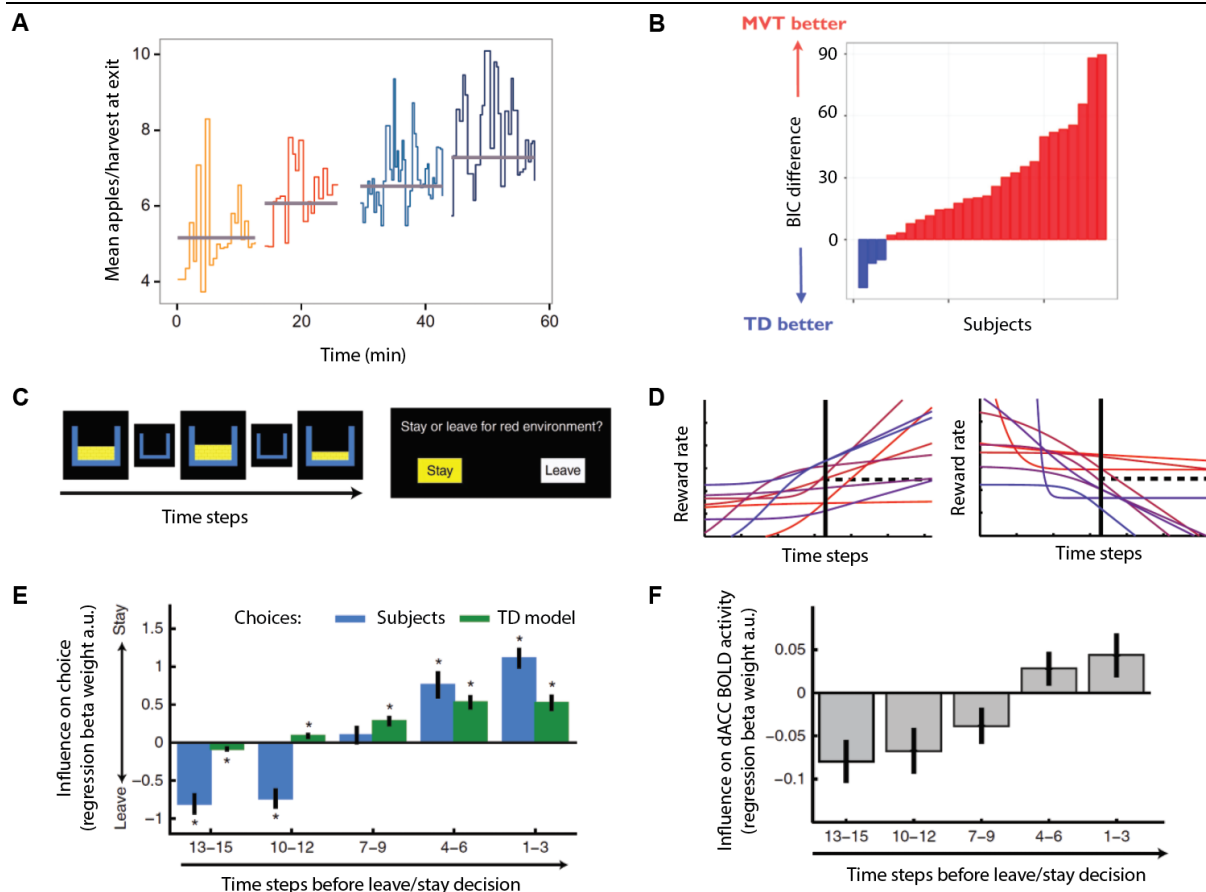


Figure 2. Learning in human foraging tasks. A-B) A virtual patch foraging task in which subjects repeatedly harvested from trees which gradually became depleted, deciding after each harvest whether to stay with the current tree or move on to the next one. The Environment changed every 15 minutes and the travel times or tree depletion rates varied across environments giving different average richness. **A)** Number of apples per harvest at tree exit for one subject across 4 environments, the grey bars show the optimal patch leaving threshold for each environment. **B)** Choice behaviour of the great majority of subjects was better fit by a Marginal Value Rule than a model-free RL model. **C-F)** A virtual patch leaving task with increasing or decreasing reward rate trajectories (**C**). Subjects sampled returns from a patch over 15 time-steps, then chose whether to stay for another 15 steps or leave for a default patch of fixed quality. **D)** Example decreasing (left panel) and increasing (right panel) reward rate trajectories (coloured lines), the solid vertical line indicates the patch leaving decision and the dashed horizontal line the reward rate of the default patch. **E)** Regression analysis of subject's choices indicated that rewards long before the leave/stay decision promoted leaving, while rewards shortly before the decision promoted staying (blue bars), consistent with estimation of the reward rate trajectory and inconsistent with a model which compared a temporal difference estimate of the reward rate in the current patch with that of the default patch (green bars) **F)** BOLD activity in dorsal anterior cingulate cortex at the decision time was negatively influence by rewards long before and positively influence by rewards shortly before the decision. **A-B** Adapted from [10], **C – F** from [13].

Constantino and Daw [10] directly addressed this question using human behaviour on a virtual foraging task. Subjects chose between harvesting a tree, obtaining apples at a cost of harvesting time, or moving on to a new tree at the cost of travel time. As subjects repeatedly harvested a tree the number of apples returned diminished. Subjects foraged in a sequence of environments which differed with respect to their travel time or tree depletion rate, giving the environments different overall richness. Subjects harvested individual trees longer in lower quality environments (Figure

2A) as predicted by MVT, and similar to the behaviour of monkeys in laboratory patch foraging tasks [6,12]. Using Bayesian model comparison, the authors asked whether trial-by-trial choices were more consistent with model-free RL algorithms or the Marginal Value Rule. The great majority of subjects were better fit by the Marginal Value Rule and support for this model was overwhelming at the group level (Figure 2B). Subjects did exhibit a tendency to over-harvest each tree compared to the optimal leaving point, which model-fitting suggest may be related to risk sensitivity.

However, another recent study on human patch leaving behaviour suggests that the Marginal Value Rule cannot be a complete picture of learning in such tasks. Wittmann et al. examined choices in a foraging task where the Marginal Value Rule was not optimal because reward rate within a patch could either increase or decrease with time [13]. Subjects repeatedly sampled a patch whose returns were stochastic but had an underlying increasing or decreasing trend (Figure 2C,D). They then chose whether to stay in the patch or switch to a default patch with a fixed, known reward rate. Rewards obtained shortly before the leave/stay decision promoted staying in the original patch, however rewards obtained long before the decision promoted patch leaving (Figure 2E). In other words, while the reward rate at the time of the leave/stay decision influenced choice, the gradient of the reward rate also influenced the decision, with subjects more likely to stay when reward rates were increasing.

The authors compared subjects choices to a model similar to the Marginal Value Rule in which a recency weighted average of the current patch's reward rate was compared with the reward rate of the default patch to decide whether to stay or leave. While this model captured the tendency of recent rewards to promote staying in the patch, it did not capture the effect of rewards long before the leave/stay decision to promote leaving (Figure. 2E), and hence the sensitivity to reward rate gradient exhibited by the subjects. The authors then considered a model that estimated the reward rate gradient by averaging reward prediction errors, with a positive gradient promoting staying in the patch and a negative gradient promoting leaving. This model better predicted subjects choices and captured the opposing influence of early and late rewards on the leave/stay decision. These data suggest that rather than simply using a smoothed retrospective estimate of the instantaneous reward rate in the patch, human subjects extrapolated the reward rate gradient to estimate the patch's future reward rate.

Model-based patch evaluation

Such prospective prediction of the future state of the patch appears characteristic of model-based RL. However model-based RL has been considered unlikely to underlie patch leaving decisions because deep and precise decision tree search would be needed to generate optimal choices [10]. One way to ameliorate planning costs is to combine model-based and model-free RL, for example by using model-based evaluation for only the next few steps and substituting model-free value estimates in place of deeper search of the decision tree [40]. Recent evidence indicates humans use such short range planning in multi-step decision tasks [36].

We suggest that model-based prediction of future returns in the current patch may improve patch leaving decisions without deep search of the decision tree. Given a model of a patch's reward rate dynamics which predicts the future reward rate within it (henceforth termed a patch model), we propose a strategy whereby the forager remains in the patch while the predicted future reward rate over any n time-steps within it is higher than the average reward rate achieved in the environment (Box 3). This strategy is motivated by two considerations:

Firstly, in environments where the forager gains noisy information about a patch's quality while foraging within it, the optimal strategy is to remain in the current patch while the expected future reward rate within the patch is higher than the average reward rate under the optimal policy [37] (Box 2). The proposed strategy converts this to a decision rule by substituting the patch model's prediction for the true expected reward rate in the patch and the average reward rate currently achieved by the forager for that under the optimal policy.

Secondly, the proposed strategy can be seen as estimating the relative values of staying and leaving by combining a model-based prediction of future returns in the patch with a model-free estimate of the quality of the broader environment (Box 3). Specifically the decision variable is an estimate of the value difference between staying and leaving under the foragers current behavioural policy in an average reward RL framework (Boxes 1,3). The strategy can therefore be seen as a form of generalised policy iteration (GPI); the process of improving a behavioural policy by iteratively making value estimates consistent with current policy, and policy greedy (choosing the highest valued actions) with respect to current value estimates [4]. Hence, we conjecture that given a good patch model and sufficient exploration, the proposed strategy will converge to a good policy.

What form might patch models take? The simplest is to use the current reward rate as a prediction of the future reward rate, in which case the proposed strategy reduces to the Marginal Value Rule. This reframes the Marginal Value Rule as a form of average reward RL algorithm which estimates relative values of staying and leaving using a particular assumption about the patch's reward rate dynamics. A simple model for changing reward rates is linear extrapolation of the current reward rate gradient. Humans readily learn linear relationships from observations and tend to assume linearity when extrapolating beyond presented data in function learning experiments [41,42]. Moreover, Wittmann et al. [13] showed that an estimate of the reward rate gradient approximated by averaging reward prediction errors was predictive of subject's choices. We therefore suggest that model-based patch evaluation is a possible common framework for understanding both the apparent prospective prediction of future reward rates in [13], and the Marginal Value Rule like behaviour in [10], though we acknowledge that the standard Marginal Value Rule and prospective patch evaluation are not readily distinguishable with the monotonically depleting patches used in [10].

In addition to the reward history, patch models may employ other evidence to better predict the future reward rate, including the statistics of previously encountered patches and non-reward cues about the quality of the current patch. However, though the input to patch models may take diverse forms, we suggest the concept is narrow enough to be testable because to qualify as such, a learning system must predict a specific quantity; the future reward rate trajectory in the patch (Box 3).

Reward rate prediction in Anterior Cingulate Cortex?

Wittmann et al. observed a possible signature of prospective prediction of the future reward rate in BOLD activity in anterior cingulate cortex (ACC) [13]. ACC activity at the time of the leave /stay decision was positively modulated by recent rewards (Figure 2F) but negatively modulated by earlier rewards, consistent with prediction of the future reward rate from the gradient of the experienced reward rate. ACC has previously been implicated in foraging decisions [6,7,9,13], but also carries signals consistent with model-based action evaluation in simultaneous choice tasks [31,43,34]. ACC neuronal populations encode reward history with a diversity of time constants and hence in principle carry information necessary for reward rate gradient estimation [44]. Furthermore, when the environment changes abruptly, network activity appears to reset in rodent ACC [45], which could be used to re-initialize reward rate tracking when leaving a patch or environment. ACC also carries

signals related to environment volatility that could change the speed of learning or horizon of integration for reward rate computations [46,47].

The other critical decision variable in both the proposed strategy and the standard Marginal Value Rule is the average reward rate in the environment. Tonic levels of dopamine in the nucleus accumbens have been proposed to encode this [48], and pharmacological manipulation of dopamine affects behavioural vigour consistent with changing the perceived opportunity cost of time [48,49]. If patch leaving decisions utilise dopaminergic average reward rate signals, boosting tonic dopamine should promote patch leaving. Though this has not to our knowledge been directly tested, pharmacologically boosting dopamine levels in a simultaneous choice task increased monkeys probability of trying novel options that were periodically introduces rather than staying with their previously preferred choice [50].

Given the ubiquity of foraging across organisms with radically different cognitive capacities, a range of mechanisms from simple heuristics to model-based evaluation are likely employed. Focussing on data from recent human patch leaving experiments, we suggest that this large brained primate learns predictive models of patch reward rate trajectories which support leave/stay decision making. Dissecting the contribution of prospective models, model-free RL, and decision heuristics is a central challenge for future work on the neurobiology of foraging.

Box 1. Average reward reinforcement learning

Most RL algorithms estimate the values of states and/or actions, defined as the long run reward they are expected to lead to. In tasks with no end point the long run rewards are unbounded, potentially leading to infinite values. The standard way of avoiding this is to exponentially discount future rewards. The value of taking action a in state s , and thereafter following behavioural policy π is then defined as the expected exponentially discounted future reward:

$$Q^\pi(s, a) = E_\pi \left\{ \sum_{t=0}^{\infty} \gamma^t r_{t+1} \middle| s_0 = s, a_0 = a \right\}$$

Where E_π is the expectation under behavioural policy π , γ is a discount parameter $0 < \gamma \leq 1$, and r_{t+1} is the reward on time-step $t+1$.

Values have recursive relationships such that the value of a given state or action can be defined in terms of the expected immediate reward, action-state transition probabilities, and values of the possible successor states or actions. These allow learning rules to incrementally improve value estimates by updating them to minimise prediction errors between the previous value estimate and the experienced immediate reward and value of the next available actions. Q-learning is one such update rule for action values:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right)$$

Where r is the immediate reward and s' the state reached after taking action a in state s . α is a learning rate which determines the size of the updates on each step.

The optimal foraging literature typically considers foragers to be maximising the reward (e.g. energy gain) per unit time rather than the exponentially discounted future reward. In the RL literature the problem of maximising reward per unit time is termed average reward RL [51]. Average reward RL has been used as a normative framework for understanding behaviours where maximising returns per unit time is the natural objective [48], including matching on concurrent interval schedules [48,52], behavioural vigour [48,53] and labour-leisure decisions [54].

Average reward RL defines values under policy π relative to the average reward per timestep ρ^π under that policy, where:

$$\rho^\pi = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n E_\pi \{r_t\}$$

Action values are then defined as:

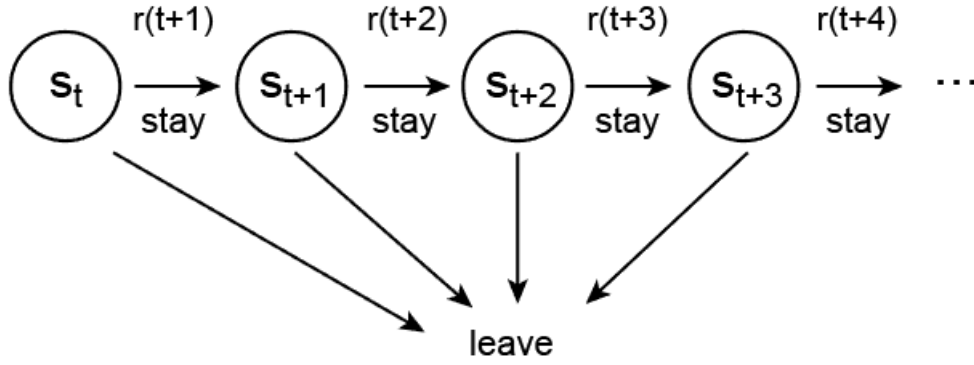
$$Q^\pi(s, a) = E_\pi \left\{ \sum_{t=0}^{\infty} r_{t+1} - \rho^\pi \middle| s_0 = s, a_0 = a \right\}$$

Values in an average reward RL framework represent the transient advantage or disadvantage of starting with a particular state or action.

This gives rise to a temporal difference learning algorithm called R-learning which is the average reward equivalent to Q-learning:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left(r - \rho + \max_{a'} Q(s', a') - Q(s, a) \right)$$

In addition to updating the action or state values, the agents estimate of the average reward rate ρ is also updated on each time-step based on experienced rewards.

Box 2. Patch leaving in average reward RL.


We consider patch leaving decisions in an average reward RL framework (Box 1). From the definition of action values in average reward RL (Box 1), the value of remaining in a patch for n time-steps, then leaving the patch and subsequently following policy π is:

$$Q^\pi(stay_n) = \sum_{t=0}^{n-1} (E_p\{r_{t+1}\} - \rho^\pi) + \sum_{t=n}^{\infty} (E_\pi\{r_{t+1}\} - \rho^\pi)$$

Where $E_p\{r_{t+1}\}$ is the expected reward rate on time-step $t + 1$ foraging in the current patch, $E_\pi\{r_{t+1}\}$ is the expected reward rate on time-step $t + 1$ having left the patch and subsequently followed policy π , ρ^π is the average reward rate under policy π . The first term is the contribution of rewards obtained in the patch over the next n time-steps and the second term is the contribution of rewards obtained after leaving the patch.

Assuming that the environment outside the current patch does not change while foraging within it (a standard assumption in MVT [1]), the second term is independent of how long the forager remains in the patch. We term it $Q^\pi(leave)$ as it is the action value for leaving the patch and subsequently following policy π .

$$Q^\pi(stay_n) = \sum_{t=0}^n (E_p\{r_{t+1}\} - \rho^\pi) + Q^\pi(leave)$$

The value difference between staying in the patch for n time-steps vs leaving immediately is:

$$Q^\pi(stay_n) - Q^\pi(leave) = \sum_{t=0}^n (E_p\{r_{t+1}\} - \rho^\pi)$$

Staying in the patch is higher valued than leaving immediately if:

$$\begin{aligned} \max_n (Q^\pi(stay_n) - Q^\pi(leave)) &> 0 \\ \max_n \left(\sum_{t=0}^n (E_p\{r_{t+1}\} - \rho^\pi) \right) &> 0 \\ \max_n \left(\frac{1}{n} \sum_{t=0}^n E_p\{r_{t+1}\} \right) &> \rho^\pi \end{aligned}$$

As the above holds for any policy including the optimal policy π^* , an optimal forager should remain in the patch while:

$$\max_n \left(\frac{1}{n} \sum_{t=0}^n E_p \{r_{t+1}\} \right) > \rho^*$$

i.e. while the expected future reward rate in the patch over any n time-steps is higher than the average reward rate ρ^* under the optimal policy.

The above decision rule is identical to that derived by McNamara [37] for optimal patch leaving in a stochastic environment. If patches are monotonically depleting such that $E_p \{r_{t+1}\} \leq E_p \{r_1\}$ for all t , the above decision rule gives Charnov's classical Marginal Value Theory [18]. Reframing these earlier results in the subsequently developed optimality framework of average reward RL more closely integrates optimal foraging with other behaviours where average reward RL has established explanatory power [48,53,54].

Box 3. Model-based patch evaluation.

Can a model of the future reward rate expected from a patch be used to improve patch leaving decisions? We propose a decision rule of remaining in the current patch while:

$$\max_n \left(\frac{1}{n} \sum_{t=0}^n E_{PM}\{r_{t+1}\} \right) > \rho_{MF}^\pi$$

Where $E_{PM}\{r_{t+1}\}$ is the patch model's prediction of the future reward rate in the patch, and ρ_{MF}^π is a model-free estimate of the average reward rate under current policy π (i.e. a recency weighted average of the experienced reward rate over a timescale spanning encounters with many patches).

The rule is motivated in part by the optimal policy for patch leaving in stochastic environments, which depends on the expected future returns in the patch $E_p\{r_{t+1}\}$ and the average reward rate under the optimal policy ρ^* [37] (Box 2). Approximating $E_p\{r_{t+1}\}$ with the patch model's prediction $E_{PM}\{r_{t+1}\}$ and ρ^* with the model-free estimate of the average reward rate under the current policy ρ_{MF}^π transforms the optimal policy into the proposed decision rule.

The rule can be viewed as making a greedy choice with respect to an estimate of values under the current policy π . Staying in the patch is the greedy (higher valued) option under policy π when (Box 2):

$$\max_n \left(\frac{1}{n} \sum_{t=0}^n E_p\{r_{t+1}\} \right) > \rho^\pi$$

Approximating $E_p\{r_{t+1}\}$ with the patch model's prediction $E_{PM}\{r_{t+1}\}$ and ρ^π with the model-free estimate ρ_{MF}^π gives the proposed decision rule.

We consider a patch model to be any neural system which predicts the future reward rate trajectory $E_p\{r_{t+1}\}$. To reach a decision the patch model must be searched over a range of time horizons to estimate:

$$\max_n \left(\frac{1}{n} \sum_{t=0}^n E_p\{r_{t+1}\} \right)$$

When a planning problem is repeatedly encountered the correct solution can often be automatised into a habitual mapping from state to action [55,22,56]. If a given patch or prey type is repeatedly encountered, one way to automatise the proposed strategy would be to cache the above quantity to avoid having to search the model over range of time horizons.

Acknowledgements

The authors thank Peter Dayan and Mark Walton for comments on an earlier version of the manuscript. NK was funded by a Junior Research Fellowship from Christchurch College, Oxford. TA by Wellcome Trust Grant 202831/Z/16/Z.

References

1. Stephens DW, Krebs JR: *Foraging theory*. Princeton University Press; 1986.
 2. Stephens DW, Brown JS, Ydenberg RC: *Foraging: behavior and ecology*. University of Chicago Press; 2007.
 3. Schultz W, Dayan P, Montague PR: **A Neural Substrate of Prediction and Reward**. *Science* 1997, **275**:1593–1599.
 4. Sutton RS, Barto AG: *Reinforcement learning: An introduction*. The MIT press; 1998.
 5. Doya K: **Reinforcement learning: Computational theory and biological mechanisms**. *HFSP J.* 2007, **1**:30–40.
 6. Hayden BY, Pearson JM, Platt ML: **Neuronal basis of sequential foraging decisions in a patchy environment**. *Nat. Neurosci.* 2011, **14**:933–939.
 7. Kolling N, Behrens TE, Mars RB, Rushworth MF: **Neural Mechanisms of Foraging**. *Science* 2012, **336**:95–98.
 8. Wikenheiser AM, Stephens DW, Redish AD: **Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task**. *Proc. Natl. Acad. Sci.* 2013, **110**:8308–8313.
 9. Blanchard TC, Hayden BY: **Neurons in Dorsal Anterior Cingulate Cortex Signal Postdecisional Variables in a Foraging Task**. *J. Neurosci.* 2014, **34**:646–655.
 - • 10. Constantino SM, Daw ND: **Learning the opportunity cost of time in a patch-foraging task**. *Cogn. Affect. Behav. Neurosci.* 2015, **15**:837–853.
- This study used Bayesian model comparison of subjects trial-by-trial choices in a patch foraging task to evaluate their learning mechanisms. The Marginal Value Rule better described their behaviour than model-free RL.
11. Calhoun AJ, Tong A, Pokala N, Fitzpatrick JAJ, Sharpee TO, Chalasani SH: **Neural Mechanisms for Evaluating Environmental Variability in *Caenorhabditis elegans***. *Neuron* 2015, **86**:428–441.
 12. Blanchard TC, Hayden BY: **Monkeys Are More Patient in a Foraging Task than in a Standard Intertemporal Choice Task**. *PLOS ONE* 2015, **10**:e0117057.
 - • 13. Wittmann MK, Kolling N, Akaishi R, Chau BKH, Brown JW, Nelissen N, Rushworth MFS: **Predictive decision making driven by multiple time-linked reward representations in the anterior cingulate cortex**. *Nat. Commun.* 2016, **7**:12327.

This study looked at human patch leaving behaviour in environments where patch reward rates either increased or decreased over time. Behaviour was consistent with prospective prediction of the future reward rate, and corresponding decision variables were observed in ACC.

14. Carter EC, Redish D: **Rats value time differently on equivalent foraging and delay-discounting tasks.** *J. Exp. Psychol. Gen.* 2016, **145**:1093–1101.
15. Passingham RE, Wise SP: *The neurobiology of the prefrontal cortex: anatomy, evolution, and the origin of insight.* Oxford University Press; 2012.
16. Pearson JM, Watson KK, Platt ML: **Decision making: the neuroethological turn.** *Neuron* 2014, **82**:950–965.
17. Kacelnik A: **Central place foraging in starlings (*Sturnus vulgaris*). I. Patch residence time.** *J. Anim. Ecol.* 1984, [no volume].
18. Charnov EL: **Optimal foraging, the marginal value theorem.** *Theor. Popul. Biol.* 1976, **9**:129–136.
19. McNamara JM, Houston AI: **Optimal foraging and learning.** *J. Theor. Biol.* 1985, **117**:231–249.
20. Daw ND, Niv Y, Dayan P: **Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control.** *Nat. Neurosci.* 2005, **8**:1704–11.
21. Balleine BW, O'Doherty JP: **Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action.** *Neuropsychopharmacology* 2009, **35**:48–69.
22. Dolan RJ, Dayan P: **Goals and Habits in the Brain.** *Neuron* 2013, **80**:312–325.
23. Balleine BW, Dickinson A: **Goal-directed instrumental action: contingency and incentive learning and their cortical substrates.** *Neuropharmacology* 1998, **37**:407–419.
24. Killcross S, Coutureau E: **Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats.** *Cereb. Cortex* 2003, **13**:400–408.
25. Yin HH, Knowlton BJ, Balleine BW: **Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning.** *Eur. J. Neurosci.* 2004, **19**:181–189.
26. Yin HH, Ostlund SB, Knowlton BJ, Balleine BW: **The role of the dorsomedial striatum in instrumental conditioning.** *Eur. J. Neurosci.* 2005, **22**:513–523.
27. Dayan P, Berridge KC: **Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation.** *Cogn. Affect. Behav. Neurosci.* 2014, **14**:473–492.
28. McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G: **Ventral Striatum and Orbitofrontal Cortex Are Both Required for Model-Based, But Not Model-Free, Reinforcement Learning.** *J. Neurosci.* 2011, **31**:2700.
29. Robinson MJF, Berridge KC: **Instant Transformation of Learned Repulsion into Motivational "Wanting."** *Curr. Biol.* 2013, **23**:282–289.

30. Sadacca BF, Jones JL, Schoenbaum G: **Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework.** *eLife* 2016, **5**:e13665.
31. 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.
32. Simon DA, Daw ND: **Neural Correlates of Forward Planning in a Spatial Decision Task in Humans.** *J. Neurosci.* 2011, **31**:5526–5539.
33. Lee SW, Shimojo S, O'Doherty JP: **Neural Computations Underlying Arbitration between Model-Based and Model-free Learning.** *Neuron* 2014, **81**:687–699.
34. Doll BB, Duncan KD, Simon DA, Shohamy D, Daw ND: **Model-based choices involve prospective neural activity.** *Nat. Neurosci.* 2015, **18**:767–772.
35. Huys QJ, Lally N, Faulkner P, Eshel N, Seifritz E, Gershman SJ, Dayan P, Roiser JP: **Interplay of approximate planning strategies.** *Proc. Natl. Acad. Sci.* 2015, **112**:3098–3103.
- • 36. Keramati M, Smittenaar P, Dolan RJ, Dayan P: **Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum.** *Proc. Natl. Acad. Sci.* 2016, **113**:12868–12873.

Using a multi-step decision task the authors find evidence that humans plan a few steps into the future then substitute model-free values for deeper search of the decision tree. Depth of planning was modulated by time pressure suggesting a speed accuracy trade-off controls planning depth.

37. McNamara J: **Optimal patch use in a stochastic environment.** *Theor. Popul. Biol.* 1982, **21**:269–288.
38. Oaten A: **Optimal foraging in patches: A case for stochasticity.** *Theor. Popul. Biol.* 1977, **12**:263–285.
39. Lima SL: **Downy Woodpecker Foraging Behavior: Efficient Sampling in Simple Stochastic Environments.** *Ecology* 1984, **65**:166–174.
40. Daw ND, Dayan P: **The algorithmic anatomy of model-based evaluation.** *Philos. Trans. R. Soc. B Biol. Sci.* 2014, **369**:20130478.
41. DeLosh EL, Busemeyer JR, McDaniel MA: **Extrapolation: The sine qua non for abstraction in function learning.** *J. Exp. Psychol. Learn. Mem. Cogn.* 1997, **23**:968.
- 42. Lucas CG, Griffiths TL, Williams JJ, Kalish ML: **A rational model of function learning.** *Psychon. Bull. Rev.* 2015, **22**:1193–1215.

Reviews theoretical models of how humans interpolate and extrapolate from observed data points to estimate underlying relationships between variables. Predicting patch reward rate trajectories from the reward history can be seen as an example of such function learning.

43. Cai X, Padoa-Schioppa C: **Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex.** *J. Neurosci.* 2012, **32**:3791–3808.
44. Bernacchia A, Seo H, Lee D, Wang X-J: **A reservoir of time constants for memory traces in cortical neurons.** *Nat. Neurosci.* 2011, **14**:366–372.

45. Karlsson MP, Tervo DG, Karpova AY: **Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty.** *Science* 2012, **338**:135–139.
46. Behrens TE., Woolrich MW, Walton ME, Rushworth MF.: **Learning the value of information in an uncertain world.** *Nat. Neurosci.* 2007, **10**:1214–1221.
47. Ligaya K: **Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system.** *eLife* 2016, **5**:e18073.
- 48. Niv Y, Daw ND, Joel D, Dayan P: **Tonic dopamine: opportunity costs and the control of response vigor.** *Psychopharmacology (Berl.)* 2007, **191**:507–520.

This paper established average reward RL as a normative framework for understanding vigour in free operant tasks, and proposed that tonic dopamine encodes the average reward rate signal.

49. Beierholm U, Guitart-Masip M, Economides M, Chowdhury R, Düzel E, Dolan R, Dayan P: **Dopamine modulates reward-related vigor.** *Neuropsychopharmacology* 2013, **38**:1495–1503.
50. Costa VD, Tran VL, Turchi J, Averbeck BB: **Dopamine modulates novelty seeking behavior during decision making.** *Behav. Neurosci.* 2014, **128**:556–566.
51. Mahadevan S: **Average reward reinforcement learning: Foundations, algorithms, and empirical results.** *Mach. Learn.* 1996, **22**:159–195.
52. Herrnstein RJ: **On the law of effect.** *J. Exp. Anal. Behav.* 1970, **13**:243–266.
53. Guitart-Masip M, Beierholm UR, Dolan R, Duzel E, Dayan P: **Vigor in the Face of Fluctuating Rates of Reward: An Experimental Examination.** *J. Cogn. Neurosci.* 2011, **23**:3933–3938.
54. Niyogi RK, Breton Y-A, Solomon RB, Conover K, Shizgal P, Dayan P: **Optimal indolence: a normative microscopic approach to work and leisure.** *J. R. Soc. Interface* 2014, **11**:20130969.
55. Dickinson A: **Actions and habits: the development of behavioural autonomy.** *Philos. Trans. R. Soc. B Biol. Sci.* 1985, **308**:67–78.
56. Akam T, Costa R, Dayan P: **Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task.** *PLoS Comput Biol* 2015, **11**:e1004648.