

Dispatch

Reinforcement Learning: Full Glass or Empty — Depends Who You Ask

Jacob J.W. Bakermans¹, Timothy H. Muller² and Timothy E.J. Behrens^{1,3}

An extension of the prediction error theory of dopamine, imported from artificial intelligence, represents the full distribution over future rewards rather than only the average and better explains dopamine responses.

The relationship between neuroscience and artificial intelligence is a long and fruitful one, with many of the greatest advances in each inspired by the other [1]. This is exemplified by the field of reinforcement learning – how, from rewards and punishments, we learn the values of our actions to choose better next time [2]. Formalising observations from animal behaviour led to simple mathematical models of how animals learnt the expected reward (or *value*) that would accrue from choosing particular stimuli, or taking particular actions [3]. Artificial intelligence researchers noticed limitations in these models and developed a general value-learning algorithm – temporal difference learning [2]. Neuroscientists then discovered that the key signal underlying temporal difference learning – the temporal difference prediction error – matched the firing of dopamine neurons with beautiful precision [4]. This work paved the way for entire fields of neuroscience, but left open important questions: While temporal difference learning learns the average – or expected – values of actions, it loses track of all the different outcomes that might follow a given action. Additionally, it did not predict the diversity of dopamine responses that is observed, suggesting not all dopamine neurons track the same expected reward. Decades after this initial discovery, a recent paper in *Nature* by

Dabney et al 2020 [5] import another insight from artificial intelligence into neuroscience that elegantly solves both these problems: The diversity of dopamine responses in fact reflects the representation of the probability distribution over different possible outcomes.

Reinforcement learning is finding how to act in order to maximise future reward. This is a difficult task because often it is not clear how actions now cause favourable outcomes later on. Once you achieve your goal, how do you know which actions were crucial for getting there? In the 1980s, artificial intelligence researchers developed the *temporal difference learning* algorithm to overcome this problem [2]. Temporal difference learning not only keeps track of reward that directly follows an action, but also the expected rewards accumulating in the extended future – when you are enjoying a hot cappuccino, you should credit value to the earlier decision to leave the office, so you are likely to make it again. Temporal difference learning updates value using the temporal difference *reward prediction error* – the difference between received and expected long-term average reward. This ensures that, after many experiences, the value of the coffee percolates back to the office door.

This temporal difference prediction error precisely explains the firing of dopamine neurons [4]. For example, the dopamine neurons of a coffee aficionado would initially fire only to the taste of the coffee. After many trips to the coffee shop, dopamine neurons start to fire upon leaving the office, in expectation of coffee in the (not-so-distant) future [4].

However, this now ‘classical’ theory of temporal difference learning only tracks *the average* – or *expectation* – of possible future reward. It would therefore struggle with problems that require knowledge (i.e. representation) of the distribution over possible outcomes of an action.

What does it mean to represent the distribution over possible outcomes? Imagine a bird that has to decide where to forage, to gather enough nutrients to make it through the night (Figure 1) [6,7]. Either it flies to its favourite berry bush, which provides a steady amount of food, or it chooses a big apple tree. The apple tree has most likely already been raided by other birds, but if not, it has a lot of juicy apples. On *average*, the berry bush yields the most fruit. But since the berries are insufficient to survive the night, the apple tree gamble is the better option. In order to realise this, the brain must represent the possibility of the two different outcomes of foraging at the apple tree. Simply averaging them, as in temporal difference learning, would result in preferring the berries and guarantee starvation.

In addition to this computational problem, dopamine neurons do not all fire equally; some respond with stronger prediction errors than others to the same reward [8,9]. This is surprising: if the neurons implement the classic temporal difference learning algorithm, they all learn the same (expected) reward and therefore signal the same prediction error. Could this diversity actually be a signature of a more complex and sophisticated mechanism underlying dopamine signalling?

Dabney et al [5] draw inspiration from artificial intelligence once again to show this diversity reflects representation of a distribution over possible outcomes. Recently, machine learning agents for reinforcement learning problems (like videogames) have been improved by *distributional reinforcement learning* [10,11]. Instead of learning a single prediction, of expected reward, a distributional reinforcement learning agent learns a whole set of different predictions, allowing it to represent the full distribution of possible rewards. Encouraged by the biological plausibility of the mechanisms for distributional reinforcement learning, the

authors set out to find its signatures in the firing of dopamine neurons – and they find compelling evidence.

The different predictions that allow a distributional reinforcement learning agent, or brain, to represent the full reward distribution can be thought of as differences in ‘optimism’ across neurons (Figure 2). An optimistic neuron predicts high reward. It will therefore produce reward prediction errors that are mostly negative, as usually the experienced reward is lower than its expectation. These neurons inform the agent about the higher end of the reward distribution. A pessimistic neuron predicts low reward, will usually produce positive prediction errors, and is informative of the lower end of the reward distribution. By contrast, in the classical temporal difference learning theory, all neurons predict the same amount of reward – the average reward. Above-average rewards always produce positive prediction errors and below-average rewards always negative prediction errors. Therefore, these two theories can be distinguished by the reversal points – the reward at which lower rewards produce negative, and higher rewards positive prediction errors – of neurons across the population.

Dabney et al [5] tested for signatures of distributional reinforcement learning using dopamine neurons previously recorded [12] in the ventral tegmental area of mice while the animals received a random amount of reward (one of 0.1, 0.3, 1.2, 2.5, 5, or 10 μ l) on each trial. Consistent with distributional reinforcement learning, the authors found a range of different reversal points across the population. Concretely, some (pessimistic) neurons produced a positive reward prediction error at 0.3 μ l, while other (optimistic) neurons produced a negative reward prediction error at 5 μ l.

What makes a neuron optimistic or pessimistic – what determines its reversal point? A single change to classical temporal difference learning leads to a diversity of reversal points.

In classical temporal difference learning, the *scaling* of reward prediction errors is identical for positive and negative reward prediction errors. Positive and negative prediction errors balance each other when they appear equally often: when the neuron predicts the average reward. Now imagine a neuron with a steeper scaling to – or higher rate of learning from – positive reward prediction errors. These strong positive reward prediction errors are balanced out when they are less frequent than weaker negative reward prediction errors. This equilibrium occurs when the neuron's reversal point, its reward prediction, is high; we obtain an optimistic neuron (Figure 2C). The opposite, a steeper scaling of negative reward prediction errors, leads to pessimistic neurons. The theory therefore predicts diversity in the slopes of the positive versus the negative prediction errors across the population of dopamine neurons, and furthermore that this should correlate with the reversal points – not expected by random diversity across neurons. Dabney et al [5] present strong evidence for both these predictions.

Distributional reinforcement learning posits that the specific set of reward predictions encodes the full distribution of rewards. The authors put this to the test in the final and perhaps most ambitious analysis (Figure 3). Remarkably, they are able to decode the probability density function of the reward distribution in the task from the set of reversal points and associated response asymmetries. They successfully reconstruct multimodal distributions from cell activity alone. Such distributions are particularly interesting because the mean is a bad approximation of the full distribution. The possible outcomes of choosing the apple tree in our example are a typical case of a multimodal distribution.

Dabney et al [5] thus provide evidence of distributional reinforcement learning in the population of dopamine neurons in the mouse ventral tegmental area. This work answers old questions and raises new. The hungry bird described above needed access to the reward distribution to make a good choice, but this is not the reason it is beneficial to artificial neural networks. In most cases, in fact, distributional reinforcement learning algorithms still make choices on the basis of the average reward. Learning the reward distribution is beneficial to these agents because it makes them build a better internal representation of their inputs. Two situations with the same average reward but different distributions would be treated as the same in classical temporal difference learning, but distributional reinforcement learning forces the network to separate them. Biological brains also have to transform sensory input into useful representations. This raises an intriguing question. Is the function of distributional reinforcement learning the same in mouse and machine, or could each benefit from the reward distribution for completely different reasons?

References:

1. Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron* 95, 245–258.
2. Sutton, R., and Barto, A. (1998). Reinforcement learning: An introduction. (MIT Press).
3. Rescorla, R., and Wagner, A. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement (Classical conditioning II: Current research and theory, 64-99).
4. Schultz, W., Dayan, P., and Montague, P.R. (1997). A Neural Substrate of Prediction and Reward. *Science*. 275, 1593–1599.

5. Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine- based reinforcement learning. *Nature* 577, 671–675.
6. Kolling, N., Wittmann, M., and Rushworth, M.F.S. (2014). Multiple neural mechanisms of decision making and their competition under changing risk pressure. *Neuron* 81, 1190–1202.
7. Caraco, T. (1981). Energy Budgets , Risk and Foraging Preferences in Dark-Eyed Juncos (*Junco hyemalis*). *Behav. Ecol. Sociobiol.*, 213–217.
8. Fiorillo, C., Tobler, P., and Schultz, W. (2003). Discrete Coding of Reward Dopamine Neurons. *Science*. 299, 1898–1903.
9. Lammel, S., Lim, B.K., and Malenka, R.C. (2014). Neuropharmacology Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* 76, 351–359.
10. Bellemare, M.G., Dabney, W., and Munos, R. (2017). A Distributional Perspective on Reinforcement Learning. *Proc. 34th Int. Conf. Mach. Learn.* 70, 449–458.
11. Dabney, W., Rowland, M., Bellemare, M.G., and Brain, G. (2018). Distributional Reinforcement Learning with Quantile Regression. *AAAI*, 2892–2901.
12. Eshel, N., Bukwich, M., Rao, V., Hemmelder, V., Tian, J., and Uchida, N. (2015). Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* 525, 243–246.

Affiliations and email addresses:

Jacob.bakermans@ndcn.ox.ac.uk, timothymuller127@gmail.com, behrens@fmrib.ox.ac.uk

[1] Wellcome Centre for Integrative Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford, John Radcliffe Hospital OX3 9DU, UK

[2] Institute of Neurology, Department of Clinical and Movement Neurosciences, University College London, London, UK

[3] Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3AR, UK

Figure legends:

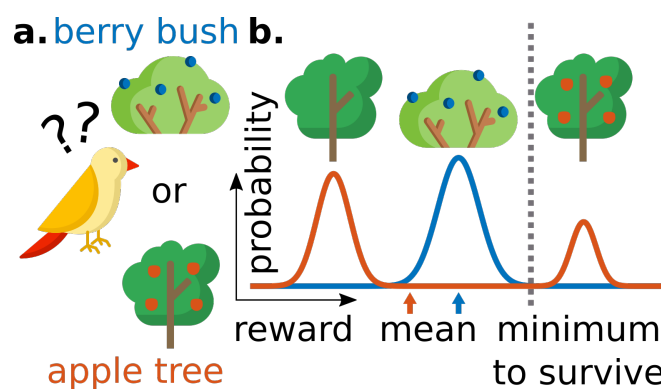


Figure 1. Surviving the night

(A) Forage at a berry bush (blue) or an apple tree (orange)? (B) The amount of food on the berry bush is higher on average (arrows), as the apple tree is likely to be depleted. But since the bird won't survive the night when its energy is below-threshold (dotted line), the apple tree is the better option. Icons by Freepik from flaticon.com.

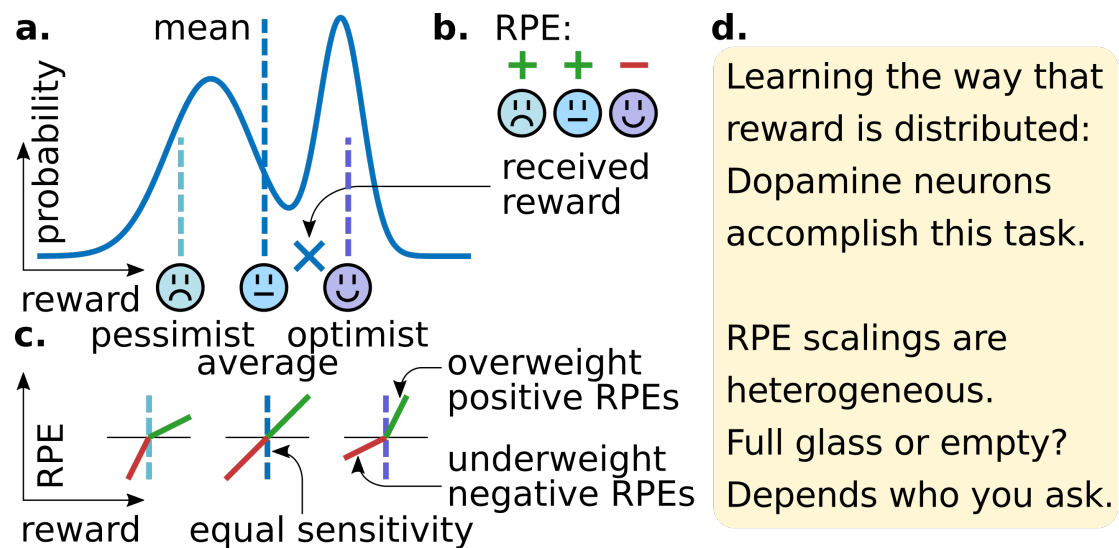


Figure 2. Learning distributions.

(A) Optimistic and pessimistic neurons predict different amounts of reward, whereas classical temporal difference neurons all predict the average reward. (B) As a consequence, distributional reinforcement learning predicts a diversity of positive and negative prediction errors for a given reward. (C) Optimistic predictions arise from the overweighting of positive reward prediction errors, and underweighting of negative ones (and vice versa for pessimistic predictions). Classical temporal difference learning neurons only predict the average reward by being equally sensitive to positive and negative reward prediction errors. (D) Summary in double dactyl.

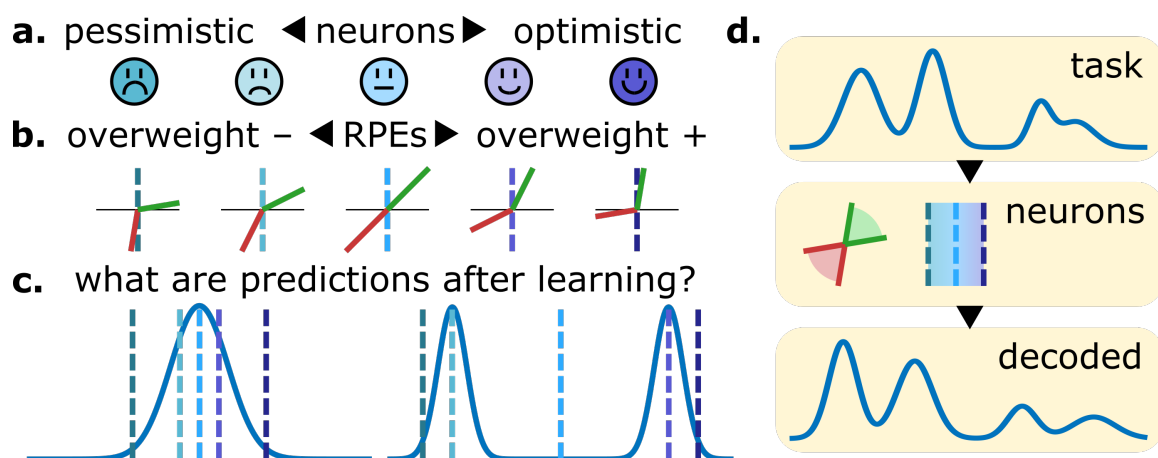


Figure 3. Reconstructing distributions

(A) Across the population, neurons are optimistic (or pessimistic) to different extents. (B) Each neuron will learn its own reward prediction (teal, blue, purple) because of differences in sensitivity to positive versus negative reward prediction errors. (C) If a very optimistic neuron learns a reward prediction of 5 μ l, reward will usually be below 5 μ l. In other words, if you know the sensitivity asymmetry of each neuron and the resulting reward prediction, you can reconstruct the underlying true distribution (like the one on the left or the right). (D) For a given task's reward distribution (top), the authors measure individual dopamine neurons (middle) and successfully decode (bottom) the ground truth.