

Psychology's History of Learning

Self-study material for DMUU Week 4

Purpose

The decision models of Weeks 1 and 2 take the value of each option as given. They do not say where that value comes from, how it was learned, or what happens when it turns out to be wrong.

This document covers the part of psychology that answered those questions, from Pavlov's salivation measurements in 1897 to the recording of a prediction-error signal in dopamine neurons in the 1990s. Until the late 1960s the working assumption was that an association forms when a stimulus and an outcome occur together often enough. An experiment called blocking (Kamin, 1969), described in Section 4, showed that this is not the case. Learning is driven by the difference between what happens and what was predicted. An outcome that is fully predicted produces no learning, however often it occurs.

This difference is the prediction error. It appears in the Rescorla–Wagner equation (1972), in the temporal difference algorithm of reinforcement learning, and in the firing of dopamine neurons. Weeks 5 and 6 build on it.

After reading this document you should be able to:

1. Distinguish Pavlovian from instrumental learning, and goal-directed from habitual control.
 2. Write down the Rescorla–Wagner equation, explain each symbol, and compute the update over several trials.
 3. Explain how the temporal difference error extends the Rescorla–Wagner error.
 4. State what the dopamine recordings show, and what they do not show.
-

How to use this document

- Read the overview table first.
- Each section describes one line of work: the question, the method, the result, and what changed.
- Section 4 contains an equation and a worked example. Do the computation yourself. Week 5 assumes you have.
- Boxes marked "Link to the Week 4 lecture" point to specific slides of the Week 4 lecture (The neuroscience of value).

- Self-check questions are at the end. They are for your own use.

Overview

Period	Figure / work	What drives learning	Method
1897–1927	Pavlov , classical conditioning	Pairing in time (contiguity)	Controlled pairing; salivation measured in drops
1898–1938	Thorndike / Skinner , instrumental learning	Consequences of action	Puzzle box; operant chamber
1930–1948	Tolman , cognitive map	Exploration, without reinforcement	Latent learning in mazes
1972	Rescorla & Wagner , formal model	Prediction error, trial by trial	A quantitative model with testable predictions
1981–1988	Sutton & Barto , temporal difference learning	Prediction error, moment by moment	Algorithm; simulation
1992–1998	Schultz; Montague, Dayan & Sejnowski	Prediction error as a neural signal	Single-unit recording in behaving monkeys; computational model

1. Pavlov: classical conditioning (1897–1927)

The question. Pavlov was a physiologist studying digestion. He measured salivation in dogs in response to food placed in the mouth. Salivation was understood as a reflex: an innate response to a specific stimulus. His dogs also salivated before food arrived, in response to stimuli that regularly preceded it. Pavlov asked whether this acquired salivation followed laws of its own (Pavlov, 1927).

The method. Pavlov controlled what the animal experienced and when. A stimulus with no prior effect on salivation was presented, followed after a fixed interval by food. The laboratory used metronomes, buzzers, lights and bells. Saliva was collected through a fistula and measured in drops. The stimulus had to produce no salivation before training. This is what makes the salivation after training unambiguously learned.

The results. Pavlov varied the arrangement systematically and described a set of phenomena that later theories had to explain:

- **Acquisition.** The response to the conditioned stimulus grows over paired trials.
- **Extinction.** The response declines when the conditioned stimulus is presented without food.
- **Spontaneous recovery.** After a rest period, part of an extinguished response returns.
- **Generalisation and discrimination.** The response transfers to similar stimuli, and can be restricted by training.
- **Conditioned inhibition.** A stimulus that signals the absence of food suppresses the response to other stimuli.

What changed. Learning became a measured quantity: drops of saliva on trial n under a specified schedule. Pavlov's proposed mechanism was contiguity: stimuli that occur close together in time become associated. Section 4 describes blocking, the experiment that showed contiguity is not sufficient.

Key terms. Conditioned stimulus (CS); unconditioned stimulus (US); conditioned response (CR); extinction.

Link to the Week 4 lecture. Slides 34–39 present Gottfried, O'Doherty and Dolan (2003). This is a Pavlovian experiment in humans in an MRI scanner. Abstract pictures are the CSs. Food odours are the USs. Amygdala and orbitofrontal responses to the pictures are CRs. Feeding the participant to satiety on one food is a devaluation test: it changes the value of the US after training and asks whether the response to the CS follows. It does.

2. Thorndike and Skinner: instrumental learning (1898–1938)

The question. In Pavlov's procedure, the US arrives whether or not the animal responds. The animal learns that one stimulus predicts another. Thorndike (1898) studied a different problem. A cat in a puzzle box receives food only if it operates the latch. The animal must learn what to do.

The method. Thorndike's puzzle box and Skinner's (1938) operant chamber make the delivery of the outcome contingent on an action. The quantity measured is the rate or latency of the action.

Two kinds of learning. The distinction survives in every modern treatment:

- **Pavlovian (classical) conditioning.** Learning which stimulus predicts which. The response is elicited, not chosen. The response repertoire is largely innate: salivation, approach, freezing, autonomic arousal.
- **Instrumental (operant) conditioning.** Learning which action produces which outcome. The response is emitted and shaped by its consequences.

Goal-directed and habitual control. Instrumental behaviour itself has two forms. The experimental test that separates them is outcome devaluation (Adams & Dickinson, 1981; Dickinson, 1985). The procedure has three steps:

1. Train a rat to press a lever for sucrose.

2. Devalue the sucrose. The rat receives sucrose paired with lithium chloride, which induces nausea. The lever is absent during this step.
3. Test lever pressing in extinction. No sucrose is delivered during the test.

The test is run in extinction so that the rat cannot learn about the new value of sucrose during the test itself. Any change in pressing must come from what the rat already represents.

After moderate training, pressing drops. The rat uses an action–outcome representation, and the outcome is no longer wanted. This is **goal-directed** control. After extended training, pressing continues. The behaviour is triggered by the stimulus and is insensitive to the current value of the outcome. This is **habitual** control.

Key terms. Pavlovian versus instrumental; goal-directed versus habitual; outcome devaluation.

Link to the Week 4 lecture. Slide 47 lists Pavlovian, habitual and goal-directed valuation systems. The Pavlovian system corresponds to Section 1. The habitual system corresponds to the behaviour that survives devaluation. The goal-directed system is the subject of Section 3. This mapping is a simplification; the three systems were identified by the devaluation logic above, not by the historical figures.

3. Tolman: the cognitive map (1930–1948)

The question. Thorndike's account treats learning as the strengthening of a response by reward. Tolman asked whether animals learn anything in the absence of reward.

The method. Tolman and Honzik (1930) ran three groups of rats through a maze once a day. Group 1 found food at the goal from day 1. Group 2 never found food. Group 3 found no food for ten days, then food from day 11 onward.

The result. Group 3's error rate fell on day 12 to the level of Group 1. Ten days of unrewarded runs had produced learning that showed up as soon as reward gave the animal a reason to use it. Tolman called this **latent learning**.

Interpretation. Tolman (1948) proposed that the rat acquires a representation of the environment, a **cognitive map**, that is independent of what is currently worth having. A second line of evidence, the shortcut experiment (Tolman, Ritchie & Kalish, 1946), is often cited. It has a known confound: a light above the goal. A meta-analysis found poor replicability across studies and little evidence for shortcutting (Duvell & Grieves, 2026). The evidence for a map rests on latent learning and on the devaluation results in Section 2.

What changed. There are two ways to compute what to do. One is to store a value for each action, learned from experience, and choose the largest. This is cheap and fast, and limited to situations already experienced. The other is to store a model of how the environment works and compute the best action from

it. This is expensive and slow, and can answer questions never faced before, including what to do when an outcome's value has just changed. The contrast between **model-free** and **model-based** control (Daw, Niv & Dayan, 2005) returns in Week 5, which covers model-free learning in detail and model-based learning qualitatively. It also explains the devaluation result: goal-directed control consults a model and detects that the outcome is no longer wanted; habitual control has only a cached value.

Key terms. Latent learning; cognitive map; model-free versus model-based control.

4. Rescorla and Wagner: a formal model of conditioning (1972)

This section has two parts. The first describes blocking, the experiment that showed pairing in time is not sufficient for learning. The second describes the model that explains it.

Blocking. By 1970, several results had shown that contiguity is not sufficient for learning. The clearest is **blocking** (Kamin, 1969). The design has three phases:

1. Phase 1: stimulus A is followed by the US until A reliably produces the CR.
2. Phase 2: A and a new stimulus B are presented together, followed by the US.
3. Test: B is presented alone. It produces little or no CR.

In a control group that skips Phase 1 and receives only the AB trials, B produces a strong CR. Both groups receive the same number of B–US pairings with the same timing. The only difference is what the animal already knew about A. Rescorla (1968) showed a related result: a CS that is followed by the US on some trials produces little conditioning if the US also occurs, at the same rate, when the CS is absent.

The conclusion is that learning depends on whether the US was already predicted. Rescorla and Wagner (1972) asked whether a single rule could generate blocking together with Pavlov's phenomena.

The model. On every trial, the associative strength of each cue present changes in proportion to the difference between the US that occurred and the US that was predicted:

$$\Delta V_A = \alpha_A \cdot \beta \cdot (\lambda - \Sigma V)$$

Symbol	Name	Meaning
V_A	associative strength of cue A	how strongly A currently predicts the US

Symbol	Name	Meaning
ΔV_A	update	the change in V_A on this trial
λ	asymptote	the maximum strength the US can support; $\lambda = 0$ when no US occurs
ΣV	prediction	the sum of V over all cues present on this trial
$\lambda - \Sigma V$	prediction error	the difference between what occurred and what was predicted
α_A	salience of cue A	learning rate associated with the cue
β	learning rate of the US	learning rate associated with the US

The update to cue A depends on ΣV , the prediction formed by all cues present, not on V_A alone. Cues compete for a fixed amount of associative strength.

A worked example. Take $\lambda = 100$ and $\alpha \cdot \beta = 0.3$ for a single cue, starting from $V = 0$.

Trial	V before	Prediction error ($\lambda - V$)	$\Delta V = 0.3 \times \text{error}$	V after
1	0	100	30.0	30.0
2	30.0	70.0	21.0	51.0
3	51.0	49.0	14.7	65.7
4	65.7	34.3	10.3	76.0

The increments shrink because the prediction error shrinks. The model therefore produces a negatively accelerated acquisition curve.

Blocking in the model. Phase 1 trains A to asymptote: $V_A = 100$. In Phase 2, $\Sigma V = V_A + V_B = 100 + 0 = 100$. The prediction error is $\lambda - \Sigma V = 0$. Therefore $\Delta V_B = 0$ on every Phase 2 trial. B acquires no associative strength. The model was built to account for this result.

Other phenomena.

- **Extinction.** The CS is presented without the US, so $\lambda = 0$. The prediction error is $0 - V$, which is negative. V declines.
- **Conditioned inhibition.** Interleave $A \rightarrow \text{US}$ trials with $AB \rightarrow \text{no US}$ trials. On the compound trials the prediction error is negative and is shared between A and B. V_B becomes negative. B predicts the absence of the US. The model gives Pavlov's conditioned inhibition a quantitative form.

- **Overexpectation.** This is the model's novel prediction (Kremer, 1978). Train $A \rightarrow US$ and $B \rightarrow US$ separately to asymptote. Then present $AB \rightarrow US$. $\Sigma V = 2\lambda$, the prediction error is negative, and both V_A and V_B decline, although every trial is reinforced. The result was confirmed.

Where the model fails. The model is quantitative and makes specific predictions. Several are wrong. Miller, Barnet and Grahame (1995) review them. Four matter for this course.

1. **Latent inhibition.** Presenting a CS alone before conditioning slows later learning about it (Lubow & Moore, 1959). In the model, V stays at 0 during pre-exposure, so pre-exposure should have no effect.
2. **Backward blocking.** Phase 1: $AB \rightarrow US$. Phase 2: $A \rightarrow US$. Test B. The model updates only cues present on a trial, so V_B should be unchanged by Phase 2. In humans, the response to B is reduced (Shanks, 1985). Two kinds of model account for it: a modified Rescorla–Wagner rule with a negative learning rate for absent cues (Van Hamme & Wasserman, 1994), and Bayesian models that track the covariance between cues (Kang, Tobler & Dayan, 2024). Week 6 introduces Bayes' rule, on which these models rest.
3. **Extinction.** The model treats extinction as a decrease in V . Renewal, reinstatement and spontaneous recovery show that the original association survives extinction and that extinction adds a second, context-dependent association (Bouton, 2004).
4. **Time.** The model updates once per trial. It has no representation of time within a trial, and no notion of rewards that lie several steps ahead. Section 5 addresses this.

Key terms. Prediction error; associative strength; learning rate; asymptote; blocking.

5. Sutton and Barto: temporal difference learning (1981–1988)

Origin. Sutton and Barto (1981) published their first model in *Psychological Review* as a theory of classical conditioning. It extended the Rescorla–Wagner model to events within a trial. The temporal difference (TD) algorithm followed (Sutton, 1988) and became a central method of reinforcement learning in computer science (Sutton & Barto, 2018).

Reading before you continue. Vandelaer (2022), *Reinforcement Learning: An introduction*, Parts 1 and 2 (Medium; PDFs on Moodle). Part 1 explains what reinforcement learning is and introduces the exploration-exploitation trade-off. Part 2 defines the terms used below: state, action, reward, policy, return, discount factor, value function. Parts 3 and 4 of the series cover a different algorithm (REINFORCE) and are not required.

The question. A trial is the experimenter's unit, not the animal's. Real USs arrive some time after real CSs. An animal that wants to act well must predict not only the next reward but the total reward still to come.

The model. Let s_t be the state at time t , r_t the reward received at time t , and γ a discount factor between 0 and 1. The **value** of a state, $V(s)$, is defined as the expected sum of future rewards from that state, each discounted by γ per step. The TD error at time t is:

$$\delta_t = r_t + \gamma \cdot V(s_{t+1}) - V(s_t)$$

and the value is updated by $V(s_t) \leftarrow V(s_t) + \alpha \cdot \delta_t$.

Compare the two errors term by term:

	Rescorla–Wagner	Temporal difference
prediction	ΣV	$V(s_t)$
target	λ , the US that occurred	$r_t + \gamma \cdot V(s_{t+1})$, the reward received now plus the discounted prediction from the next state
error	$\lambda - \Sigma V$	$r_t + \gamma \cdot V(s_{t+1}) - V(s_t)$
time step	one trial	one moment within a trial

The structure is the same. The target has changed. TD learning corrects a prediction against the next prediction, not only against the outcome. This is called **bootstrapping**.

A consequence. Because each state's value is corrected against the next state's value, the error propagates backward through a sequence over repeated trials. The error moves from the time of the reward to the time of the earliest reliable predictor of the reward.

What changed. TD learning is an algorithm. It runs on problems no animal has been tested on. It also makes a claim about a signal rather than about behaviour: if a brain implements TD learning, some quantity in it should behave like δ_t . That quantity should respond positively to an unpredicted reward, shift its response to a predictive cue after training, and go negative at the moment a predicted reward fails to arrive.

For the curious mind (optional, on Moodle). Two papers show reinforcement learning outside the laboratory. Jayaraman et al. (2024) introduce RL to clinicians and discuss its use for sequential treatment decisions, including the difficulty of defining a reward for a patient. Naeem, Rizvi and Coronato (2020), Sections II–IV, give a broader tutorial on the algorithms, including SARSA, which Week 5 covers.

Key terms. Temporal difference error; value function; discount factor; bootstrapping.

6. Schultz: dopamine neurons and the prediction error (1992–1998)

The recordings. Schultz and colleagues recorded single dopamine neurons in the midbrain of monkeys during simple conditioning tasks (Schultz, Apicella & Ljungberg, 1993). A dopamine neuron fires a few spikes per second even when nothing happens. This is its baseline rate. The recordings compare firing against this baseline in three conditions:

1. **Unpredicted reward.** Juice is delivered without warning. The neuron fires a burst: firing rises well above baseline for a fraction of a second.
2. **Predicted reward.** After training with a cue, the burst at the juice disappears. The neuron fires at the cue instead.
3. **Predicted reward omitted.** After training, the cue is presented but no juice follows. The neuron fires at the cue. At the time the juice was due, the neuron briefly stops firing: its rate falls below baseline.

Condition 2 shows that the neuron does not respond to reward as such. Condition 3 shows that the neuron encodes when a reward was expected. Nothing happens at the moment the firing falls. A change in firing at that moment is only possible if the neuron's input carries a prediction of when the juice should arrive.

The interpretation. Montague, Dayan and Sejnowski (1996) showed that the three conditions match the behaviour of the TD error δ_t . Schultz, Dayan and Montague (1997) presented the correspondence to a general audience. The order was: recordings first, interpretation second. One prediction of the TD account was tested afterwards: when the timing of reward changes, the dopamine response tracks the error in the predicted time (Hollerman & Schultz, 1998).

What the recordings do not show. The reward prediction error (RPE) account is well supported for a large population of midbrain dopamine neurons. It is not the only function of dopamine circuits. Gershman et al. (2024) review four qualifications:

1. The phasic burst has an early component driven by stimulus intensity and salience that precedes the value-specific component.
2. Some dopamine subpopulations respond to aversive as well as rewarding events, and different subpopulations project to different striatal targets.
3. Dopamine neurons respond to errors in the predicted sensory identity of a reward even when its value is unchanged. A model-free value signal does not predict this.
4. Dopamine release covaries with movement. An average-reward account predicts the link to vigor. The link to movement direction and kinematics is not predicted by a value account.

Key terms. Reward prediction error (RPE).

Learning rate and uncertainty

In the Rescorla–Wagner model the learning rate is a constant. Humans adjust it. When the environment changes often, people weight recent evidence more heavily; when it is stable, they weight it less (Behrens et al., 2007). Bayesian models derive the learning rate from the learner's current uncertainty (Kang, Tobler & Dayan, 2024). Week 6 introduces Bayes' rule, on which these models rest.

Where each topic returns

From this document	Returns in
Blocking; prediction error	Week 5
Rescorla–Wagner equation	Week 5, assumed known
Temporal difference learning	Week 5
Dopamine as an RPE signal, with its qualifications	Week 5
Cognitive map; model-free and model-based control	Week 5
Learning rate as a function of uncertainty	Week 6 (Bayes' rule)

Self-check questions

For your own use. If you can answer these, you are ready for Week 5. The section that covers each question is given in brackets.

1. **(Section 2.)** A rat is trained to press a lever for sucrose. The sucrose is then devalued: the rat receives it paired with nausea, without the lever present. At test, the lever is available but no sucrose is delivered. Does the rat press less than before? Answer separately for moderate and for extended training, and name the control system in each case.
2. **(Section 4, blocking.)** In the blocking design, cue B is paired with the US as often as cue A was. Why does B acquire no associative strength? Answer in terms of ΣV .
3. **(Section 4, worked example.)** Using $\lambda = 100$ and $\alpha\beta = 0.2$, compute V for the first three trials starting from zero. Is learning faster or slower than in the worked example? Which symbol is responsible?

4. **(Section 4, extinction.)** A cue has $V = 100$. It is now presented repeatedly without the US. State λ for these trials, compute the prediction error on the first trial, and say what happens to V .
 5. **(Section 3.)** In Tolman and Honzik (1930), Group 3 received no food in the maze for ten days and food from day 11. Describe what happened on day 12 and state what it shows about the role of reinforcement in learning.
 6. **(Section 6.)** In the dopamine recordings, the response to juice disappears once a cue predicts it (condition 2, "predicted reward"). Explain why this is evidence against dopamine as a reward signal. Use the term "prediction error."
 7. **(Section 6.)** In condition 3 ("predicted reward omitted"), the cue is presented but no juice follows. At the time the juice was due, the neuron's firing falls below its baseline rate. What does this show that condition 2 does not?
 8. **(Section "Learning rate and uncertainty".)** The Rescorla–Wagner model has a fixed learning rate. Describe one environment in which this is the wrong choice, and say what should determine the learning rate instead.
 9. **(Section 4, "Where the model fails", backward blocking.)** Phase 1: $AB \rightarrow US$. Phase 2: $A \rightarrow US$. Test B. What does the Rescorla–Wagner model predict? What is observed in humans? Name one model that accounts for the observation.
 10. **(Sections 4 and 5.)** Write the Rescorla–Wagner prediction error and the TD error side by side. Identify the term that differs and explain what the difference allows the TD model to do.
-

Further reading

Primary sources

- Pavlov, I. P. (1927). *Conditioned Reflexes* (G. V. Anrep, Trans.). Oxford University Press.
- Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review, Monograph Supplements*, 2(4), i–109. <https://doi.org/10.1037/h0092987>
- Skinner, B. F. (1938). *The Behavior of Organisms*. Appleton-Century.
- Tolman, E. C., & Honzik, C. H. (1930). Introduction and removal of reward, and maze performance in rats. *University of California Publications in Psychology*, 4, 257–275.
- Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning. I. Orientation and the short-cut. *Journal of Experimental Psychology*, 36(1), 13–24.
- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, 55(4), 189–208.

- Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and Aversive Behavior* (pp. 279–296). Appleton-Century-Crofts.
- Rescorla, R. A. (1968). Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative and Physiological Psychology*, 66(1), 1–5.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton-Century-Crofts.
- Kremer, E. F. (1978). The Rescorla–Wagner model: Losses in associative strength in compound conditioned stimuli. *Journal of Experimental Psychology: Animal Behavior Processes*, 4, 22–36.
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 135–170.
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1), 9–44.
- Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13(3), 900–913.
- Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16(5), 1936–1947.
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599.
- Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience*, 1(4), 304–309.

Instrumental learning and devaluation

- Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology*, 33B(2), 109–121.
- Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society B*, 308(1135), 67–78.
- Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science*, 301(5636), 1104–1107.

On the Rescorla–Wagner model and its limits

- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla–Wagner model. *Psychological Bulletin*, 117(3), 363–386.
- Soto, F. A., Vogel, E. H., Uribe-Bahamonde, Y. E., & Perez, O. D. (2023). Why is the Rescorla–Wagner model so influential? *Neurobiology of Learning and Memory*, 204, 107794.

- Lubow, R. E., & Moore, A. U. (1959). Latent inhibition: The effect of nonreinforced pre-exposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology*, 52(4), 415–419.
- Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *Quarterly Journal of Experimental Psychology*, 37B(1), 1–21.
- Van Hamme, L. J., & Wasserman, E. A. (1994). Cue competition in causality judgments: The role of nonpresentation of compound stimulus elements. *Learning and Motivation*, 25(2), 127–151.
- Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, 11(5), 485–494.

On the cognitive map

- Duvelle, É., & Griebes, R. M. (2026). Tolman's sunburst maze 80 years on: A meta-analysis reveals poor replicability and little evidence for shortcutting. *European Journal of Neuroscience*, 63(1), e70365. <https://doi.org/10.1111/ejn.70365>

Reinforcement learning and the bridge to Weeks 5 and 6

- Vandelaer, C. (2022, August 20). Reinforcement learning: An introduction (Part 1/4). *Medium*. <https://medium.com/@cedric.vandelaer/reinforcement-learning-an-introduction-part-1-4-866695deb4d1>
- Vandelaer, C. (2022, August 20). Reinforcement learning: An introduction (Part 2/4). *Medium*. <https://medium.com/@cedric.vandelaer/reinforcement-learning-an-introduction-part-2-4-46a1491a2451>
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press. Chapters 1 and 6.
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711.
- Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9), 1214–1221.
- Kang, P., Tobler, P. N., & Dayan, P. (2024). Bayesian reinforcement learning: A basic overview. *Neurobiology of Learning and Memory*, 211, 107924.

For the curious mind (optional)

- Jayaraman, P., Desman, J., Sabounchi, M., Nadkarni, G. N., & Sakhuja, A. (2024). A primer on reinforcement learning in medicine for clinicians. *npj Digital Medicine*, 7, 337. <https://doi.org/10.1038/s41746-024-01316-0>
- Naeem, M., Rizvi, S. T. H., & Coronato, A. (2020). A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access*, 8, 209320–209344. <https://doi.org/10.1109/ACCESS.2020.3038605>

On dopamine beyond the prediction error

- Gershman, S. J., Assad, J. A., Datta, S. R., Linderman, S. W., Sabatini, B. L., Uchida, N., & Wilbrecht, L. (2024). Explaining dopamine through prediction errors and beyond. *Nature Neuroscience*, 27, 1645–1655.
-