MIT Press • Open Encyclopedia of Cognitive Science

Reinforcement Learning

Anne GE Collins[1]

[1]Department of Psychology, Helen Wills Neuroscience Institute, University of California, Berkeley

MIT Press

Published on: Dec 03, 2024

DOI: https://doi.org/10.21428/e2759450.36d1ca92

Reinforcement learning (RL) refers to a process in which an agent (biological or artificial) learns how to behave in its environment by using a simple type of information: reinforcers, which index how good or bad something is. The term RL is used by multiple scientific communities to cover different but overlapping concepts. In cognitive science, RL typically describes how animals, including humans, learn to make choices that maximize rewards and minimize punishments in the aggregate. In artificial intelligence, RL refers to a class of learning environments and algorithms that agents can use to solve them, optimizing their long-term rewards, often expressed as expected cumulative future discounted rewards. In neuroscience, the RL circuit describes a specific brain network that integrates reward information to influence future choices beneficially. Cognitive RL behavior is often modeled with RL algorithms and is dependent on the brain's RL circuit. However, other non-RL algorithms and non-RL neural processes are also essential to explaining how animals, in particular humans, learn efficiently from rewards and punishments. RL in the context of cognitive sciences is best considered as a mixture of multiple processes above and beyond RL in neuroscience and RL in artificial intelligence.

History

The roots of RL trace back to experimental psychology in the late 19th and early 20th centuries. Researchers demonstrated that both human and nonhuman animals could learn to predict reward outcomes and even complex behaviors through rewards and punishments without explicit instructions or examples. Pavlov's experiments, which showed that animals could learn to associate a ringing bell with food, laid the groundwork for understanding classical conditioning. Pioneers like Skinner and Thorndike successfully taught cats, pigeons, and other animals tasks, such as unlocking boxes and playing games, in a process termed instrumental conditioning, using only reinforcers (Skinner, 1963; Thorndike, 1913). Conditioning was at the heart of a movement in psychological research called behaviorism, in which psychologists thought that any behavior could be trained through what amounts to trial and error. These ideas about trainability included high-level human behavior and were the theoretical basis for behavioral therapy, notably advanced by Mary Cover Jones in the early 20th century (Jones, 1924).

Subsequent research into the behavior of biological agents learning through reinforcement, or *RL behavior*, revealed many interesting phenomena (Dickinson & Mackintosh, 1978). For example, learning did not always necessitate the direct experience of a reward but could be trained by a signal itself associated with a reward (such as the bell); this phenomenon, called *chaining*, was essential in understanding how complex, multi-action behaviors arise through RL. Researchers also observed that rewards only drove learning when they could not be predicted, a phenomenon called *overshadowing*. Early computational modeling efforts aimed to explain mathematically how such learning occurred and to develop simple algorithms that could capture the breadth of conditioning phenomena. For example, Rescorla and Wagner developed one of the earliest RL models that captured many (but not all) known aspects of classical conditioning (Wagner & Rescorla, 1972). The Rescorla–

Wagner model aimed to predict expected outcomes in a given state and used the difference between predicted and obtained outcomes as a teaching signal to update estimates, a subsequently common approach in RL.

Concurrently, researchers in applied mathematics developed reinforcement learning algorithms for artificial agents based on the theoretical framework of *Markov Decision Processes* (MDPs; Bellman, 1957). Over the second half of the 20th century, many types of RL algorithms were developed to improve the ability of artificial agents to learn efficiently (Sutton & Barto, 2018), and some of those algorithms shared important features with models developed by psychologists. For example, similar to the Rescorla–Wagner model, algorithms such as *Q-learning* and *Temporal Difference* (or TD) *learning* attempt to estimate the value of different states or actions (the *Q*-value in Q-learning) and update this estimate after each choice with a *reward prediction error*, computed by comparing the prediction from its prior estimate to the observed outcome (the temporal difference in TD refers to the difference in estimated value between subsequent time points; see below for a more precise definition).

The two domains converged toward the end of the 20th century when the relevance of formal RL algorithms to RL in biological agents was discovered. Notably, early work in nonhuman primates, later confirmed in many species including rodents and humans, demonstrated that dopamine—a neurotransmitter with broad projections across brain areas—appeared to signal a reward prediction error in a way well captured by TD-learning RL algorithms (Montague et al., 1996; Schultz et al., 1997). Subsequent work showed that dopamine played a causal role in promoting plasticity in associations between cortex and striatum, a subcortical region that was important for decision-making and whose activity related to choice values and strategies. This uncovered a well-defined network of brain regions that approximated an RL algorithm's implementation and thus linked biological, cognitive, and computational perspectives of RL (Doya, 2007; Niv, 2009).

The study of RL in psychology and neuroscience has since exploded, seeking deeper insights into the nature of the underlying mechanisms but also validating the usefulness of concepts originating in the behaviorist era to post-cognitive revolution psychology (Collins, 2019). Modern cognitive scientists attempt to understand how internal representations shape reinforcement learning behavior. For example, Tolman in the 1950's showed that rats built maps that helped them more quickly learn how to navigate a maze towards a reward (Tolman, 1948). Recent research in RL attempts to understand how similar principles guide highly flexible human learning (Russek et al., 2017; Whittington et al., 2022).

RL in computational domains has exploded in parallel with the deep neural network revolution since the 2010's. Deep-RL approaches apply standard RL principles of developing good policies by optimizing the same objective of expected future discounted rewards but use deep neural networks to parameterize internal representations of states, actions, value estimations, and/or policies (Van Hasselt et al., 2016). This approach has yielded tremendous progress in artificial agents learning from outcomes in both virtual and robotic environments.

Core concepts

States, actions, and rewards: MDPs

RL can operate within the theoretical framework of MDPs (Bellman, 1957). In an MDP, an agent at time t is defined as being in a given state s_t, choosing actions a_t to interact with its environment, which returns a reinforcement signal r_t according to a reward function and the agent's new state s_{t+1} according to a transition function. States encompass all the information relevant to a problem, such as the bell ringing for Pavlov's dog, the geographical position in a maze for Tolman's rats or a navigating artificial intelligence (AI), or a stimulus presented on a screen for a human performing an RL experiment. Actions can be physical/motor movements, like pressing a lever or navigating one step, or more abstract, such as choosing between two items, irrespective of the specific actions that will accomplish that selection. From a computational perspective, the reward function is a simple scalar function that is designed to define the objective of an agent (maximizing it) and so is a direct function of the goal. For biological agents, reinforcers can be primary (signals that are innately aversive or appetitive such as sucrose, water, pain, some social signals, etc.), secondary (signals that have been trained to be associated with subsequent primary reinforcers: money, tokens, points, etc.), or more abstract (gaining information, reaching a goal without external reward, etc.; Daniel & Pollmann, 2014). The transition function is a characteristic of the learning problem/environment the agent is in and indicates how the agent's state changes in response to actions.

Value estimation and policy

The objective of an RL agent is to optimize a specific objective, the expected sum of future discounted rewards:

$$\mathbf{E}\left[\sum_{\mathbf{t=0}}^{\infty}\gamma^{\mathbf{t}}\mathbf{r_t}\right] = \mathbf{r_0} + \gamma\mathbf{r_1} + \gamma^{\mathbf{2}}\mathbf{r_2} + \cdots$$

This construct measures how much cumulated reward r_t I can expect in total from this time point on (the sum from time $t = 0$ to infinity), discounting rewards in the future exponentially more than those I receive immediately (with a discount factor $0 \leq \gamma \leq 1$). The expectation *E[]* in this equation takes into account uncertainty in my future choices and in the environment.

To pursue the goal of optimizing their future rewards, RL agents can attempt to estimate these values conditioned on their current state—*V(s)*—and their potential next action—*Q(s,a)*—using the information they obtain by interacting with the environment (Sutton & Barto, 2018). These value estimates can then be used to build *policies*, which can be thought of as mappings between states and actions. For example, a greedy policy selects whichever action has the highest estimated *Q*-value in the current state. Policies are often probabilistic: for example, you may go to your favorite restaurant 90% of the time but to others 10% of the time. Alternatively, some RL algorithms focus directly on optimizing policies without explicitly estimating values. In

general, RL algorithms are designed to provide some theoretical guarantees that learned values and/or policies are a good solution to a given class of problems with respect to the objective, under some assumptions.

Exploration vs. exploitation

One essential assumption is often that the agent's policy is not greedy but includes exploration. RL presents a unique dilemma compared to other types of learning, such as supervised learning: choices are opportunities to obtain rewards but also to gain information about the problem. Especially early in learning, it may be important that agents make choices that appear worse in order to gather useful information that might lead to discovering better long-term policies. For example, a human might choose to venture further than the convenient restaurant next door and try several bad ones before they discover a great and inexpensive one two blocks away. This exploration comes at the cost of short-term rewards but enables better long-term policies to be discovered. The exploration/exploitation dilemma is an important aspect of RL behavior (Gershman, 2018) and algorithms (Kaelbling et al., 1996).

Prediction errors, model-free RL algorithms, and the brain

One of the best-known quantities of RL is the reward prediction error (RPE). RPEs, often labeled δ (delta), compute the difference between new and previous value estimation after observing information. In its simplest form, when the information is simply the reward obtained (not the next state), δ is simply $\delta = r - V$ (or $r - Q$), which compares the reward obtained, r, and the reward expected, V or Q. The simplest forms of RL algorithms (often called delta rule) use this RPE to incrementally update value estimates: $V_{t+1} \leftarrow V_t + \alpha\delta$, in which the learning rate α (alpha) controls the time constant of integrating the outcome information. Such algorithms successfully capture many aspects of learning (for example, simple stimulus–outcome associations) in humans and animals (Collins, 2019; Daw & Tobler, 2014) and converge to provably good solutions under some assumptions (Sutton & Barto, 2018).

A slightly more sophisticated version of the RPE, called the *temporal difference* (TD) RPE, takes into account not just the reward but also the next state to compare previous and new estimates (Tesauro, 1995). The new estimate is $r_t + \gamma V(s_{t+1})$, adding to the obtained reward the discounted value of the next step. Thus, the RPE is $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$. The update is the same, and this becomes a TD RL algorithm. Similar principles can be applied to state-action values leading to other algorithms, such as SARSA or Q-learning, which are crucial for environments in which actions directly influence subsequent states (Watkins & Dayan, 1992).

RL algorithms that incrementally track values with RPE updates have been especially popular in neuroscience and cognitive science because they capture many aspects of how biological agents, including humans, learn from rewards, but even more so because they have been found to have some mechanistic validity with respect to biological processes (Frank et al., 2004) [see Neuroplasticity]. Dopamine neuron firing and release appears to correlate with TD RPE in a way that causally supports learning, whereas neurons in the striatum (a

subcortical brain structure highly conserved across species) appear to support value or policy representation and to causally support reward-based decision-making and learning (Daw & Tobler, 2014; Doya, 2007; Niv, 2009).

Cognitive RL mechanisms

The RL algorithms described so far belong to a category called *model-free RL* algorithms. Such algorithms are limited in a specific way: they only track an integrated, "cached" value of what has been experienced in the past, with no easy way to rapidly adjust value estimates and policies if the environment changes. By contrast, other approaches store and use information differently: agents might have access to a *"model"* of their environment, such as knowledge of which states and actions lead to rewards (the reward function) or how taking a given action might impact the next state the agent finds itself in (the transition function). For example, one might know how good three restaurants are and what the route is to go to each. Note that this is not always known—for example, the restaurant might be new, in which case there would be uncertainty on the reward function, or the restaurant might be only accessible by bus, and there might be uncertainty as to whether the bus is running today or not. When agents have a model of the environment, they can use it to estimate the value of states and actions or to design a policy in a way that is not only dependent on past experience but also on the model. Such algorithms are called model-based (as opposed to model-free RL like TD or Q-learning; Sutton & Barto, 2018).

Using a model can make agents more flexible and, in that sense, is an important approach to modeling human cognition. Such a model use can, for example, be proactive in model-based planning: agents can simulate forward trajectories in their problem and use them to decide, without relying on past experience (e.g., information that restaurant one is a $2 bus ticket, $30 cost, 8/10 delicious and that restaurant two is next door, $50, 10/10 delicious can lead to estimated subjective values; Doll et al., 2012). Models can also be used in other ways such as model-based inference—for example, to help identify the state the agent is in (Doya et al., 2002). As an example, if you know that the restaurant has two chefs on different days with different best dishes but do not know which one is cooking today, you may use your model after the first course to infer today's chef (delicious onion soup, chef A) and consequently decide accordingly (avoid the crème brûlée for dessert) and learn (chef A also makes excellent ratatouille).

Questions, controversies, and new developments

RL research is often cited as a success story of computational cognitive neuroscience, as it is a rare example in which cross-domain pollination has been tremendously successful, with computational algorithms of RL leading to novel insights in neuroscience and psychology alike, and in which insights from cognition and brain sciences have also inspired AI research. This synergy between computational algorithms and cognition has enriched both fields, yet RL remains an evolving area with unresolved questions, controversies, and continuous advancements.

One controversy that crosses fields is the following question: Is RL enough? In the context of AI, the corresponding debate centers on whether RL algorithms can universally solve learning problems by appropriately framing them within RL paradigms. Given the recent tremendous progress in AI RL algorithms, this is an important question. Some leading AI researchers advocate for this universality (Silver et al., 2021), although the usefulness of this framing is a controversial suggestion (Abel et al., 2021) [see The Free Energy Principle].

In cognition and neuroscience, this question is also relevant, in a slightly narrower interpretation. Many very diverse forms of learning can be well described by reinforcement learning algorithms, including the extreme examples of a very slow acquisition of motor skills (Fu & Anderson, 2006) or implicit associations (Cortese et al., 2020) and a few shot learning of rules (Collins & Frank, 2012). Furthermore, there are reasons to think that RL-like computations in the brain can support more complex cognitive functions than are typically considered in the context of RL, by applying it to different state, action, or reward in parallel, for example (Collins, 2018; Hazy et al., 2006). However, recent research also challenges this “omnipresence” of RL in multiple ways by highlighting how other non-RL processes may be confused for RL processes (Gershman & Daw, 2017; Yoo & Collins, 2022). Specifically, it is known that the brain has multiple independent but interactive memory mechanisms, and the flexibility of RL algorithms can lead to misinterpreting contributions of such mechanisms (such as working memory) for RL processes. An important question for future cognition research is to clarify what RL can and cannot explain in biological behavior, focusing on not only algorithms but also interpretable processes (Eckstein et al., 2021).

One way in which RL’s applicability is broadened is by reframing any problem with the appropriate reward function such that behavior optimizes it (Silver et al., 2021). This opens the question of the definition of RL problems in the context of biological animals—while RL research has often focused on the algorithm itself (given the state, action, reward, and transition function, how is a policy or value estimation learned?), a fundamental yet often overlooked question is “What are the states and actions?” (Niv, 2019; Rmus et al., 2021) and, additionally, what is the reward function for animals’ RL processes (Karayanni & Nelken, 2022; McDougle et al., 2022)? Given an algorithm and environment, different state, action, and reward definitions are certain to lead to extremely different behavior. However, this is typically taken for granted in cognition RL literature and predefined in RL AI literature, although notable exceptions exist in both fields (Karayanni & Nelken, 2022; Singh et al., 2010). Future research must consider how internal representations of states, actions, and rewards are conceptualized and utilized, enhancing our understanding of RL from both the "wet" (biological) and "dry" (computational) perspectives.

In the context of cognition, another controversy surrounds how to parse out the processes that support learning. There is broad agreement that the decision-making relies on multiple separable processes. In particular, a dichotomy between habitual (rigid, automatic, and effortless) and goal-directed (flexible, outcome-sensitive, and effortful) behaviors is well recognized (Daw & Dayan, 2014). Yet, mapping these high-level descriptions

of behavior to specific separable neural mechanisms and computational processes remains challenging and controversial. For example, the goal-directed vs. habitual dichotomy is often mapped to the computational RL notion of model-based vs. model-free learning, in which MB-RL consists of using a known model of the transitions and rewards in the environment to estimate a good policy by planning it forward, for example, through dynamic programming. However, there is increasing evidence that this approach does not adequately parse out relevant underlying cognitive processes in a way that is interpretable in terms of neural substrates and translatable to broad categories of learning (Miller et al., 2018).

More generally, standard RL algorithms (including model-based planning) often fall short of capturing the breadth of human learning. Other models, sometimes inspired by successful AI approaches, sometimes by adjacent cognitive domains, are also shown to be relevant explananda for behavior and brain function. Indeed, flexible learning also recruits different cognitive processes (Rmus et al., 2021). For example, humans may use working memory to explicitly remember specific aspects of their policy (sort baby's clothes on the left bin, kids on the right; Yoo & Collins, 2022). They may also use specific events in long-term episodic memory to guide choices, by identifying their similarity to a current state (Gershman & Daw, 2017). Progress in parsing out the processes that support flexible learning in humans will require further careful investigation of different cognitive mechanisms and their interactions (Collins & Cockburn, 2020). Such findings in the cognitive domain should inform AI research, leading to the development of innovative (non-RL) algorithms that support more flexible RL behavior (Whittington et al., 2020).

Recent developments also challenge our understanding of how RL is implemented in the brain. While evidence for the role of dopamine as a reward prediction-like teaching signal for cortico-striatal plasticity supporting reward-optimizing policies remains strong, our understanding has also become more nuanced as more complex patterns have emerged (Berke, 2018). For example, it is becoming increasingly clear that dopamine signaling is richer and less homogeneous than previously thought, integrating more than reward and encoding more than RPEs in various contexts and pathways. There are also many questions regarding how we learn to approach reward vs. avoid bad outcomes and whether a unidimensional scalar value (as is natural in algorithms) is indeed how biological agents learn. For example, researchers have explored the role of other neurotransmitters such as serotonin in avoidance learning. Acetylcholine and norepinephrine have also been theorized to play important roles in regulating learning, for example, by adjusting the rate of learning to the changeability of the environment, or to help with credit assignment (Doya, 2002)

Furthermore, even accepting that there is a well-defined brain network that implements model-free RL-like computations, the different cognitive processes that support flexible learning rely on different brain networks (Rmus et al., 2021). Explaining human RL requires not only considering non-RL algorithms as well as non-RL brain processes. Thus, research investigating the broader questions of how the brain represents and uses knowledge of the environment to learn more efficiently are important new developments. For example, regions outside of the typical RL brain network, such as the hippocampus and orbitofrontal cortex, appear to play a role

in the representation of cognitive maps (Wilson et al., 2014), which may support flexible types of learning algorithms beyond more simple RL ones.

Broader connections

RL frameworks have emerged as a cornerstone of the recent AI/deep learning surge, tackling challenges previously deemed exclusive to human intelligence such as mastering the game of Go or navigating complex multiplayer first-person games (Silver et al., 2017). At the same time, as deep RL shows itself to be increasingly successful at solving many problems, including robotic applications in the physical world (Finn et al., 2017), it does not do so in an efficient way compared with much more sample-efficient human learning, who reach comparable levels of performance with order of magnitude fewer attempts (Lake et al., 2017).

Contrary to RL AI agents that typically start their training from a blank slate, humans may benefit from many *inductive biases*: preexisting knowledge about their environment that can be used to learn more efficiently [see Bayesianism]. In addition, advances in AI considering *lifelong learning* (Abel et al., 2018) or *meta-learning*, in which agents learn to behave in many different environments non-independently, are likely to provide important steps toward making RL agents less training-sample greedy (Duan et al., 2016). In an extreme example, meta-RL (or RL2) frameworks use very slow RL algorithms to train the weights of a recurrent networks; the trained (fixed-weights) network then behaves in a fast RL-like behavior without explicitly implementing an RL algorithm, enabling the agent to adapt much faster to a novel environment, and potentially mimicking non-RL processes of the brain (such as working memory). Meta-cognition, in general, may be crucial for successful RL behavior: an agent may need to adjust how it learns in different environment contexts (for example, learning fast in fast changing environments but more incrementally in stable but noisy environments). Meta-cognitive processes may allow agents to tune their internal algorithms to behave adaptively.

RL also has an important relationship to the field of behavioral economics, which concerns itself with how biological agents make choices based on uncertain stakes, exemplified with lotteries. Risk seeking/risk aversion, loss aversion, serial dependencies in choice, context dependencies, and other apparently suboptimal biases have been revisited in the context of learning (i.e., in which values and probabilities are learned through experience, rather than instructed) and sometimes been shown to be (resource-)rational reflections of the properties of our environment (Palminteri & Lebreton, 2022).

Finally, RL is highly relevant to the more recent fields of computational psychiatry/quantitative clinical experimental psychology, a subpart of which seeks to use cognitive modeling to bridge between populations' clinical symptoms and underlying mechanisms in a way that is interpretable and translatable towards treatment (Gueguen et al., 2021). In that sense, better understanding which neurotransmitter plays which role in learning, whether working memory vs. RL processes are responsible for learning impairments in a clinical population or

whether orbitofrontal cortex vs. hippocampus helps hold cognitive maps for planning, could have far-reaching implications for future treatment avenues.

As reinforcement learning continues to evolve, its integration with cognitive science, neuroscience, and other disciplines not only enhances our understanding of artificial intelligence but also enriches our insights into human cognition. Reinforcement learning, therefore, remains a vibrant field of study with potential impacts far beyond its original confines, promising to contribute significantly to both theoretical knowledge and practical applications (Radulescu et al., 2019).

Further reading

- RL and the brain: Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005
- Broadening RL: Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8
- RL from AI perspective: Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT Press.

References

- Abel, D., Dabney, W., Harutyunyan, A., Ho, M. K., Littman, M., Precup, D., & Singh, S. (2021). On the expressivity of Markov reward. *Advances in Neural Information Processing Systems*, *34*, 7799–7812. ↩
- Abel, D., Jinnai, Y., Guo, S. Y., Konidaris, G., & Littman, M. (2018). Policy and value transfer in lifelong reinforcement learning. *International Conference on Machine Learning*, 80, 20–29. ↩
- Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, *6*, 679–684. https://doi.org/10.1512/IUMJ.1957.6.56038 ↩
- Berke, J. D. (2018). What does dopamine mean? *Nature Neuroscience*, *21*(6), 787–793. https://doi.org/10.1038/s41593-018-0152-y ↩
- Collins, A. G. E. (2018). Learning structures through reinforcement. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making* (pp. 105–123). Elsevier. ↩

↩

- Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576–586. https://doi.org/10.1038/s41583-020-0355-6

↩

- Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*(7), 1024–1035. https://doi.org/10.1111/j.1460-9568.2011.07980.x

↩

- Cortese, A., Lau, H., & Kawato, M. (2020). Unconscious reinforcement learning of hidden brain states supported by confidence. *Nature Communications*, *11*(1), 4429. https://doi.org/10.1038/s41467-020-17828-8

↩

- Daniel, R., & Pollmann, S. (2014). A universal role of the ventral striatum in reward-based learning: Evidence from human studies. *Neurobiology of Learning and Memory*, *114*, 90–100. https://doi.org/10.1016/j.nlm.2014.05.002

↩

- Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478. https://doi.org/10.1098/rstb.2013.0478

↩

- Daw, N. D., & Tobler, P. N. (2014). Value learning through reinforcement: The basics of dopamine and reinforcement learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics* (pp. 283–298). Elsevier.

↩

- Dickinson, A., & Mackintosh, N. (1978). Classical conditioning in animals. *Annual Review of Psychology*, *29*(1), 587–612. https://doi.org/10.1146/annurev.ps.29.020178.003103

↩

- Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*(6), 1075–1081. https://doi.org/10.1016/j.conb.2012.08.003

↩

- Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4-6), 495–506. https://doi.org/10.1016/s0893-6080(02)00044-8

↩

-

↩

- Doya, K., Samejima, K., Katagiri, K.-i., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369. https://doi.org/10.1162/089976602753712972

↩

- Duan, Y., Schulman, J., Chen, X., Bartlett, P. L., Sutskever, I., & Abbeel, P. (2016). RL^2: Fast reinforcement learning via slow reinforcement learning. *arXiv*. https://doi.org/10.48550/arXiv.1611.02779

↩

- Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? Interpreting model parameters in cognition and neuroscience. *Current Opinion in Behavioral Sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

↩

- Finn, C., Yu, T., Zhang, T., Abbeel, P., & Levine, S. (2017). One-shot visual imitation learning via meta-learning. *arXiv*. https://doi.org/10.48550/arXiv.1709.04905

↩

- Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, *306*(5703), 1940–1943. https://doi.org/10.1126/science.1102941

↩

- Fu, W.-T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*(2), 184. https://doi.org/10.1037/0096-3445.135.2.184

↩

- Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

↩

- Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128. https://doi.org/10.1146/annurev-psych-122414-033625

↩

- Gueguen, M. C., Schweitzer, E. M., & Konova, A. B. (2021). Computational theory-driven studies of reinforcement learning and decision-making in addiction: What have we learned? *Current Opinion in Behavioral Sciences*, *38*, 40–48. https://doi.org/10.1016/j.cobeha.2020.08.007

↩

- Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2006). Banishing the homunculus: Making working memory work. *Neuroscience*, *139*(1), 105–118. https://doi.org/10.1016/j.neuroscience.2005.04.067 ↩
- Jones, M. C. (1924). A laboratory study of fear: The case of Peter. *Pedagogical Seminary*, *31*(4), 308–315. https://doi.org/10.1080/00221325.1991.9914707 ↩
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*(1), 237–285. ↩
- Karayanni, M., & Nelken, I. (2022). Extrinsic rewards, intrinsic rewards, and non-optimal behavior. *Journal of Computational Neuroscience*, *50*(2), 139–143. https://doi.org/10.1007/s10827-022-00813-z ↩
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837 ↩
- McDougle, S. D., Ballard, I. C., Baribault, B., Bishop, S. J., & Collins, A. G. (2022). Executive function assigns value to novel goal-congruent outcomes. *Cerebral Cortex*, *32*(1), 231–247. https://doi.org/10.1093/cercor/bhab205 ↩
- Miller, K. J., Ludvig, E. A., Pezzulo, G., & Shenhav, A. (2018). Realigning models of habitual and goal-directed decision-making. In R. Morris, A. Bornstein, & A. Shenhav (Eds.), *Goal-directed decision making* (pp. 407–428). Elsevier. ↩
- 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. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996 ↩
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. https://doi.org/10.1016/j.jmp.2008.12.005 ↩
- Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8

↩

- Palminteri, S., & Lebreton, M. (2022). The computational roots of positivity and confirmation biases in reinforcement learning. *Trends in Cognitive Sciences*, *26*(7), 607–621. https://doi.org/10.1016/j.tics.2022.04.005

↩

- Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: The role of structure and attention. *Trends in Cognitive Sciences*, *23*(4), 278–292. https://doi.org/10.1016/j.tics.2019.01.010

↩

- Rmus, M., McDougle, S. D., & Collins, A. G. (2021). The role of executive function in shaping reinforcement learning. *Current Opinion in Behavioral Sciences*, *38*, 66–73. https://doi.org/10.1016/j.cobeha.2020.10.003

↩

- Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J., & Daw, N. D. (2017). Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Computational Biology*, *13*(9), e1005768. https://doi.org/10.1371/journal.pcbi.1005768

↩

- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

↩

- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driescche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270

↩

- Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. https://doi.org/10.1016/j.artint.2021.103535

↩

- Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*(2), 70–82. https://doi.org/10.1109/TAMD.2010.2051031

↩

-

↩

- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

↩

- Tesauro, G. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM*, *38*(3), 58–68. https://doi.org/10.1145/203330.203343

↩

- Thorndike, E. L. (1913). *The psychology of learning* (Vol. 2). Teachers College, Columbia University.

↩

- Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189. https://doi.org/10.1037/h0061626

↩

- Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. *arXiv*. https://doi.org/10.48550/arXiv.1509.06461

↩

- Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In *Inhibition and learning* (pp. 301–336). Erlbaum.

↩

- Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292. https://doi.org/10.1007/BF00992698

↩

- Whittington, J. C., McCaffary, D., Bakermans, J. J., & Behrens, T. E. (2022). How to build a cognitive map. *Nature Neuroscience*, *25*(10), 1257–1272. https://doi.org/10.1038/s41593-022-01153-y

↩

- Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman-Eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, *183*(5), 1249–1263. https://doi.org/10.1016/j.cell.2020.10.024

↩

- Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. https://doi.org/10.1016/j.neuron.2013.11.005

↩

-

↩