Trends in Cognitive Sciences

CellPress
REVIEWS

Review

Reinforcement Learning, Fast and Slow

Matthew Botvinick,[1,2,*] Sam Ritter,[1,3] Jane X. Wang,[1] Zeb Kurth-Nelson,[1,2] Charles Blundell,[1] and Demis Hassabis[1,2]

Deep reinforcement learning (RL) methods have driven impressive advances in artificial intelligence in recent years, exceeding human performance in domains ranging from Atari to Go to no-limit poker. This progress has drawn the attention of cognitive scientists interested in understanding human learning. However, the concern has been raised that deep RL may be too sample-inefficient – that is, it may simply be too slow – to provide a plausible model of how humans learn. In the present review, we counter this critique by describing recently developed techniques that allow deep RL to operate more nimbly, solving problems much more quickly than previous methods. Although these techniques were developed in an AI context, we propose that they may have rich implications for psychology and neuroscience. A key insight, arising from these AI methods, concerns the fundamental connection between fast RL and slower, more incremental forms of learning.

Powerful but Slow: The First Wave of Deep RL

Over just the past few years, revolutionary advances have occurred in artificial intelligence (AI) research, where a resurgence in neural network or 'deep learning' methods [1,2] has fueled breakthroughs in image understanding [3,4], natural language processing [5,6], and many other areas. These developments have attracted growing interest from psychologists, psycholinguists, and neuroscientists, curious about whether developments in AI might suggest new hypotheses concerning human cognition and brain function [7–11].

One area of AI research that appears particularly inviting from this perspective is deep RL (Box 1). Deep RL marries **neural network** modeling (see Glossary) with reinforcement learning, a set of methods for learning from rewards and punishments rather than from more explicit instruction [12]. After decades as an aspirational rather than practical idea, deep RL has within the past 5 years exploded into one of the most intense areas of AI research, generating super-human performance in tasks from video games [13] to poker [14], multiplayer contests [15], and complex board games, including go and chess [16–19].

Beyond its inherent interest as an AI topic, deep RL would appear to hold special interest for psychology and neuroscience. The mechanisms that drive learning in deep RL were originally inspired by animal conditioning research [20] and are believed to relate closely to neural mechanisms for reward-based learning centering on dopamine [21]. At the same time, deep RL leverages neural networks to learn powerful representations that support generalization and transfer, key abilities of biological brains. Given these connections, deep RL would appear to offer a rich source of ideas and hypotheses for researchers interested in human and animal learning, both at the behavioral and neuroscientific levels. And indeed, researchers have started to take notice [7,8].

At the same time, commentary on the first wave of deep RL research has also sounded a note of caution. On first blush it appears that deep RL systems learn in a fashion quite different from

Highlights

Recent AI research has given rise to powerful techniques for deep reinforcement learning. In their combination of representation learning with reward-driven behavior, deep reinforcement learning would appear to have inherent interest for psychology and neuroscience.

One reservation has been that deep reinforcement learning procedures demand large amounts of training data, suggesting that these algorithms may differ fundamentally from those underlying human learning.

While this concern applies to the initial wave of deep RL techniques, subsequent AI work has established methods that allow deep RL systems to learn more quickly and efficiently. Two particularly interesting and promising techniques center, respectively, on episodic memory and meta-learning.

Alongside their interest as AI techniques, deep RL methods leveraging episodic memory and meta-learning have direct and interesting implications for psychology and neuroscience. One subtle but critically important insight which these techniques bring into focus is the fundamental connection between fast and slow forms of learning.

[1]DeepMind, London, UK
[2]University College London, London, UK
[3]Princeton University, Princeton, NJ, USA

*Correspondence:
botvinick@google.com (M. Botvinick).

humans. The hallmark of this difference, it has been argued, lies in the sample efficiency of human learning versus deep RL. Sample efficiency refers to the amount of data required for a learning system to attain any chosen target level of performance. On this measure, the initial wave of deep RL systems indeed appear drastically different from human learners. To attain expert human-level performance on tasks such as Atari video games or chess, deep RL systems have required many orders of magnitude more training data than human experts themselves [22]. In short, deep RL, at least in its initial incarnation, appears much too slow to offer a plausible model for human learning. Or so the argument has gone [23,24].

The critique is indeed applicable to the first wave of deep RL methods, reported beginning around 2013 (e.g., [25]). However, even in the short time since then, important innovations have occurred in deep RL research, which show how the sample efficiency of deep RL can be dramatically increased. These methods mitigate the original demands made by deep RL for huge amounts of training data, effectively allowing deep RL to be fast. The emergence of these computational techniques revives deep RL as a candidate model of human learning and a source of insight for psychology and neuroscience.

In the present review, we consider two key deep RL methods that mitigate the sample efficiency problem: episodic deep RL and meta-RL. We examine how these techniques enable fast deep RL and consider their potential implications for psychology and neuroscience.

Sources of Slowness in Deep RL

A key starting point for considering techniques for fast RL is to examine why initial methods for deep RL were in fact so slow. Here, we describe two of the primary sources of sample inefficiency. At the end of the paper, we will circle back to examine how the constellations of issues described by these two concepts are in fact connected.

The first source of slowness in deep RL is the requirement for *incremental parameter adjustment*. Initial deep RL methods (which are still very widely used in AI research) employed gradient descent to sculpt the connectivity of a **deep neural network** mapping from perceptual inputs to action outputs (Box 1). As has been widely discussed not only in

Glossary

Deep neural network: a neural network with one or (typically) more hidden layers.
Embedding: a learned representation residing in a layer of a neural network.
Hidden layer: a layer of a neural network in between the input and output layers.
Neural network: a learnable set of weights and biases arranged in layers which process an input in order to produce an output. To learn more, see McClelland and Rumelhart's seminal primer [105].
Non-parametric: in a non-parametric model, the number of parameters is not fixed and can grow as more data is provided to the model.
Recurrent neural network: a neural network that runs at each time step in a sequence, passing its hidden layer activations from each step to the next.

Box 1. Deep Reinforcement Learning

RL centers on the problem of learning a behavioral policy, a mapping from states or situations to actions, which maximizes cumulative long-term reward [12]. In simple settings, the policy can be represented as a look-up table, listing the appropriate action for any state. In richer environments, however, this kind of simple listing is infeasible, and the policy must instead be encoded implicitly as a parameterized function. Pioneering work in the 1990s showed how this function could be approximated using a multilayer (or deep) neural network ([78], L.J. Lin, PhD Thesis, Carnegie Melon University, 1993), allowing gradient-descent learning to discover rich, nonlinear mappings from perceptual inputs to actions (see panel A and below). However, technical challenges prevented the integration of deep neural networks with RL until 2015, when breakthrough work demonstrated how deep RL could be made to work in complex domains such as Atari video games [13] (see Figure IB and below). Since then, rapid progress has been made toward improving and scaling deep RL [79], allowing its application to complex task domains such as Go [16] and Capture the Flag [80]. In many cases, the later advances have involved integrating deep RL with architectural and algorithmic complements, such as tree search [16] or slot-based, episodic-like memory [52] (see Figure IC and below). Other developments have focused on the goal of learning speed, allowing deep RL to make progress based on just a few observations, as reviewed in the main text.

The figure illustrates the evolution of deep RL methods, starting in panel A with Tesauro's groundbreaking backgammon-playing system 'TD-gammon' [78]. This centered on a neural network which took as input a representation of the board and learned to output an estimate of the 'state value,' defined as expected cumulative future rewards, here equal simply to the estimated probability of eventually winning a game from the current position. Panel B shows the Atari-playing DQN network reported by Mnih and colleagues [13]. Here, a convolutional neural network (see [3]) takes screen pixels as input and learns to output joystick actions. Panel C shows a schematic representation of a state-of-the art deep RL system reported by Wayne and colleagues [51]. A full description of the detailed 'wiring' of this RL agent is beyond the scope of the present paper (but can be found in [51]). However, as the figure indicates, the architecture comprises multiple modules, including a neural network that leverages an episodic-like memory to predict upcoming events, which 'speaks' to a reinforcement-learning module that selects actions based on the predictor module's current state. The system learns, among other tasks, to perform goal-directed navigation in maze-like environments, as shown in Figure I.

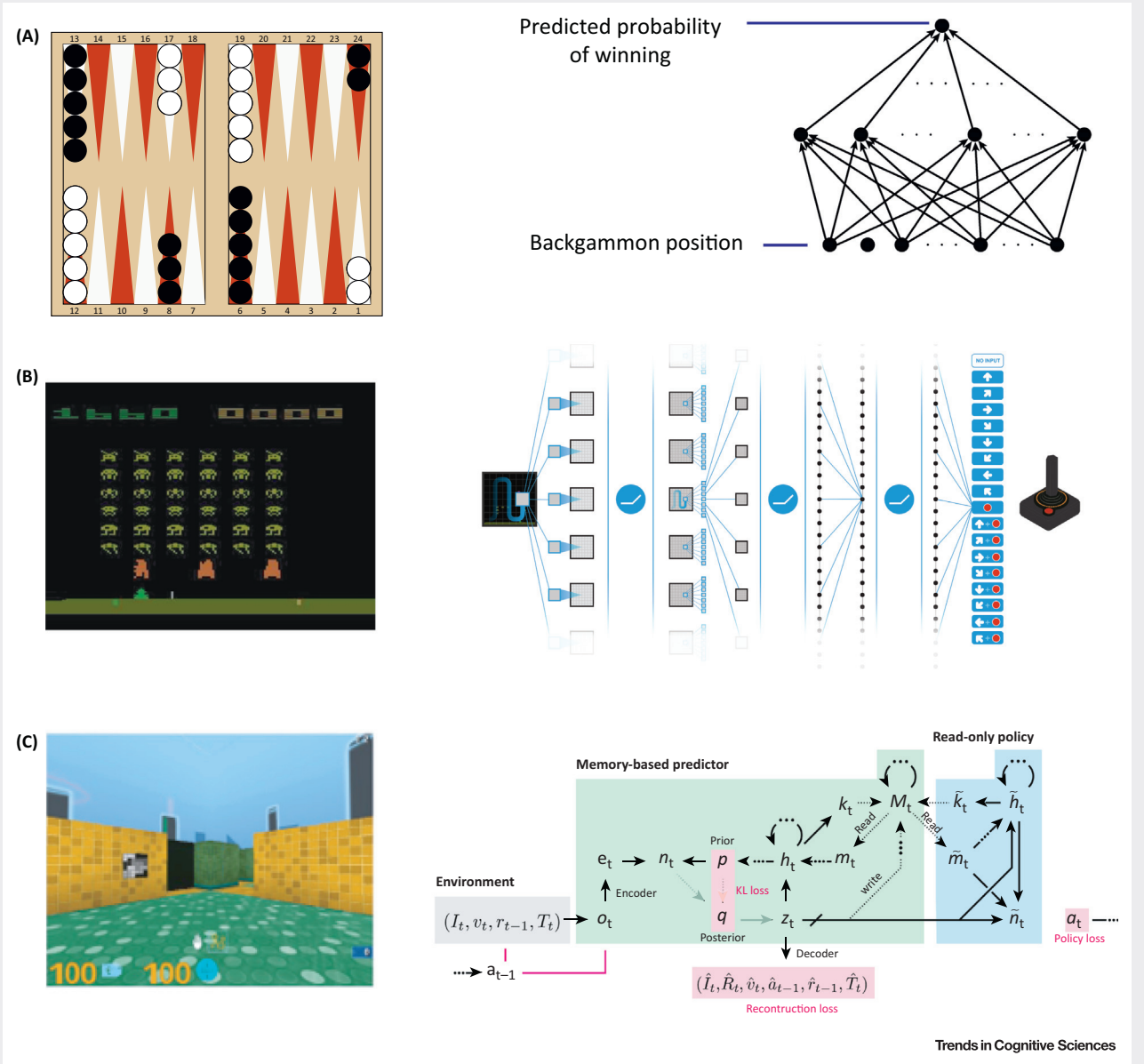

Figure I. Representative Examples of Deep Reinforcement Learning.

AI but also in psychology [26], the adjustments made during this form of learning must be small, in order to maximize generalization [27] and avoid overwriting the effects of earlier learning (an effect sometimes referred to as 'catastrophic interference'). This demand for small step-sizes in learning is one source of slowness in the methods originally proposed for deep RL.

A second source is *weak inductive bias*. A basic lesson of learning theory is that any learning procedure necessarily faces a bias–variance trade-off: the stronger the initial assumptions the learning procedure makes about the patterns to be learned (i.e., the stronger the initial inductive

bias of the learning procedure) the less data will be required for learning to be accomplished (assuming the initial inductive bias matches what's in the data!). A learning procedure with weak inductive bias will be able to master a wider range of patterns (greater variance), but will in general be less sample-efficient [28]. In effect, strong inductive bias is what allows fast learning. A learning system that only considers a narrow range of hypotheses when interpreting incoming data will, perforce, hone in on the correct hypothesis more rapidly than a system with weaker inductive biases (again, assuming the correct hypothesis falls within that narrow initial range). Importantly, generic neural networks are extremely low-bias learning systems; they have many parameters (connection weights) and are capable of using these to fit a wide range of data. As dictated by the bias–variance trade-off, this means that neural networks, in the generic form employed in the first deep RL models (Box 1) tend to be sample-inefficient, requiring large amounts of data to learn.

Together, these two factors—incremental parameter adjustment and weak inductive bias—explain the slowness of first-generation deep RL models. However, subsequent research has made clear that both of these factors can be mitigated, allowing deep RL to proceed in a much more sample-efficient manner. In what follows, we consider two specific techniques, one of which confronts the incremental parameter adjustment problem, and the other of which confronts the problem of weak inductive bias. In addition to their implications within the AI field, both of these AI techniques bear suggestive links with psychology and neuroscience, as we shall detail.

Episodic Deep RL: Fast Learning through Episodic Memory

If incremental parameter adjustment is one source of slowness in deep RL, then one way to learn faster might be to avoid such incremental updating. Naively increasing the learning rate governing gradient descent optimization leads to the problem of catastrophic interference. However, recent research shows that there is another way to accomplish the same goal, which is to keep an explicit record of past events, and use this record directly as a point of reference in making new decisions. This idea, referred to as episodic RL [29,30,42], parallels '**non-parametric**' approaches in machine learning [28] and resembles 'instance-' or 'exemplar-based' theories of learning in psychology [31,32]. When a new situation is encountered and a decision must be made concerning what action to take, the procedure is to compare an internal representation of the current situation with stored representations of past situations. The action chosen is then the one associated with the highest value, based on the outcomes of the past situations that are most similar to the present. When the internal state representation is computed by a multilayer neural network, we refer to the resulting algorithm as 'episodic deep RL'. A more detailed explanation of the mechanics of episodic deep RL is presented in Box 2.

In episodic deep RL, unlike the standard incremental approach, the information gained through each experienced event can be leveraged immediately to guide behavior. However, whereas episodic deep RL is able to go 'fast' where earlier methods for deep RL went 'slow,' there is a twist to this story: the fast learning of episodic deep RL depends critically on slow incremental learning. This is the gradual learning of the connection weights that allows the system to form useful internal representations or **embeddings** of each new observation. The format of these representations is itself learned through experience, using the same kind of incremental parameter updating that forms the backbone of standard deep RL. Ultimately, the speed of episodic deep RL is enabled by this slower form of learning. That is, fast learning is enabled by slow learning.

This dependence of fast learning on slow learning is no coincidence. As we will argue below, it is a fundamental principle, applicable to psychology and neuroscience no less than AI. Before

Box 2. Episodic Deep RL

Episodic RL algorithms estimate the value of actions and states using episodic memories [30,43,44]. Consider, for example, the episodic valuation algorithm depicted in Figure I, wherein the agent stores each encountered state along with the discounted sum of rewards obtained during the next *n* time steps. These two stored items comprise an episodic memory of the encountered state and the reward that followed. To estimate the value of a new state, the agent computes a sum of the stored discounted rewards, weighted by the similarity (sim.) between stored states and the new state. This algorithm can be extended to estimate action values by recording the actions taken along with the states and reward sums in the memory store, then querying the store to find only memories in which the to-be-evaluated action was taken. In fact, [81] used such an episodic RL algorithm to achieve strong performance on Atari games.

The success of episodic RL depends on the state representations used to compute state similarity. In a follow-up to [81], Pritzel *et al.* [29] showed that performance could be improved by gradually shaping these state representations using gradient descent learning. These results demonstrated strong performance and state of the art data efficiency on the 57 games in the Atari Learning Environment [29], showcasing the benefits of combining slow (representation) learning and fast (value) learning.

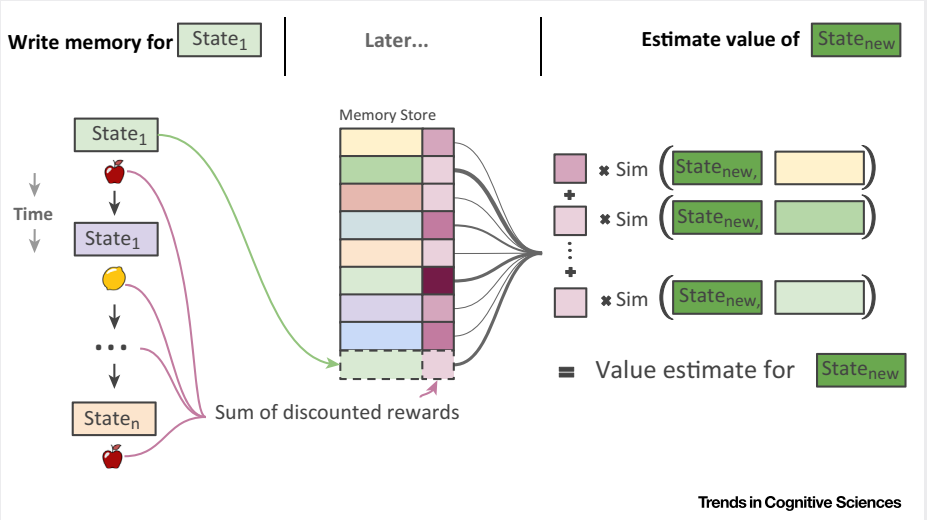

Figure I. Illustration of an Example of an Episodic Reinforcement Learning Algorithm.

turning to a consideration of this general point, however, we examine its role in the second recently developed AI technique for rapid deep RL: meta-RL.

Meta-RL: Speeding up Deep RL by Learning to Learn

As discussed earlier, a second key source of slowness in standard deep RL, alongside incremental updating, is weak inductive bias. As formalized in the idea of the bias–variance tradeoff, fast learning requires the learner to go in with a reasonably sized set of hypotheses concerning the structure of the patterns that it will face. The narrower the hypothesis set, the faster learning can be. However, as foreshadowed earlier, there is a catch: a narrow hypothesis set will only speed learning if it contains the correct hypothesis. While strong inductive biases can accelerate learning, they will only do so if the specific biases the learner adopts happen to fit with the material to be learned. As a result of this, a new learning problem arises: how can the learner know what inductive biases to adopt?

One natural answer to this question is to draw on past experience. Of course, this self-evidently occurs all the time in daily life. Consider, for example, the task of learning to use a new smart-

phone. In this context, one's past experience with smart-phones and other related devices will inform one's hypotheses concerning the way the new phone should work and will guide exploration of the phone's operation. These initial hypotheses correspond to the 'bias' in the bias–variance tradeoff, and they are responsible for the ability to quickly learn how to use the new phone. A learner without these biases (i.e., with higher 'variance') would consider a wider range of hypotheses about the phone's operation, at the expense of learning speed.

The leveraging of past experience to accelerate new learning is referred to in machine learning as meta-learning [33]. However, not surprisingly, the idea originates from psychology, where it has been called 'learning to learn.' In the first paper to use this term, Harlow [34] presented an experiment that captures the principle neatly. Here, monkeys were presented with two unfamiliar objects, and permitted to grab one of them. Beneath lay either a food reward or an empty well. The objects were then placed before the animal again, possibly left–right reversed, and the procedure was repeated for a total of six rounds. Two new and unfamiliar objects were then substituted in, and another six trials ensued with these objects. Then another pair of objects, and so forth. Across many object pairs, the animals were able to figure out that a simple rule always held: one object yielded food and the other did not, regardless of left–right position. When presented with a new pair of objects, Harlow's monkeys were able to learn in one shot which the preferable object was, a simple but vivid example of learning to learn (Box 3).

Returning now to AI, recent work has shown how learning to learn can be leveraged to speed up learning in deep RL. This general idea has been implemented in a variety of ways [35,36]. However, one approach that has particular relevance to neuroscience and psychology was proposed simultaneously by Wang [37] and Duan [38] and their colleagues. Here, a **recurrent neural network** is trained on a series of interrelated RL tasks. The weights in the network are adjusted very slowly, so they can absorb what is common across tasks, but cannot change fast enough to support the solution of any single task. In this setting, something rather remarkable occurs. The activity dynamics of the recurrent network come to implement their own separate RL algorithm, which 'takes responsibility' for quickly solving each new task, based on knowledge accrued from past tasks (Figure 1). Effectively, one RL algorithm gives birth to another, and hence the moniker 'meta-RL'.

As with episodic deep RL, meta-RL again involves an intimate connection between fast and slow learning. The connections in the recurrent network are updated slowly across tasks, allowing general principles that span tasks to be 'built into' the dynamics of the recurrent network. The resulting network dynamics implement a new learning algorithm which can solve new problems quickly, because they have been endowed with useful inductive biases by the underlying process of slow learning (Box 3). Once again, fast learning arises from, and is enabled by, slow learning.

Episodic Meta-RL

Importantly, the two techniques we have discussed above are not mutually exclusive. Indeed, recent work has explored an approach to integrating meta-learning and episodic control, capitalizing on their complementary benefits [39,40]. In episodic meta-RL, meta-learning occurs within a recurrent neural network, as described in the previous section and Box 3. However, superimposed on this is an episodic memory system, the role of which is to reinstate patterns of activity in the recurrent network. As in episodic deep RL, the episodic memory catalogues a set of past events, which can be queried based on the current context. However, rather than linking contexts with value estimates, episodic meta-RL links them with stored activity patterns from the recurrent network's internal or hidden units. These patterns are important because, through meta-RL, they come to summarize what the agent has learned

from interacting with individual tasks (see Box 3 for details). In episodic meta-RL, when the agent encounters a situation that appears similar to one encountered in the past, it reinstates the hidden activations from the previous encounter, allowing previously learned information to immediately influence the current policy. In effect, episodic memory allows the system to recognize previously encountered tasks, retrieving stored solutions.

Through simulation work in bandit and navigation tasks, Ritter *et al.* [39] showed that episodic meta-RL, just like 'vanilla' meta-RL, learns strong inductive biases that enable it to rapidly solve novel tasks. More importantly, when presented with a previously encountered task, episodic meta-RL immediately retrieves and reinstates the solution it previously discovered, avoiding the need to re-explore. On the first encounter with a new task, the system benefits from the rapidity of meta-RL; on the second and later encounters, it benefits from the one-shot learning ability conferred by episodic control.

Implications for Neuroscience and Psychology

As we discussed at the outset, the problem of sample inefficiency has been adduced as a reason for questioning the relevance of deep RL to learning in humans and other animals [23,24]. One important implication of episodic deep RL and meta-RL, from the point of view of

Box 3. Meta-Reinforcement Learning

Meta-learning occurs when one learning system progressively adjusts the operation of a second learning system, such that the latter operates with increasing speed and efficiency [82–86]. This scenario is often described in terms of two 'loops' of learning, an 'outer loop' that uses its experiences over many task contexts to gradually adjust parameters that govern the operation of an 'inner loop', so that the inner loop can adjust rapidly to new tasks (see Figure 1). Meta-RL refers to the case where both the inner and outer loop implement RL algorithms, learning from reward outcomes and optimizing toward behaviors that yield maximal reward.

As discussed in the main text, one implementation of this idea employs a form of recurrent neural network commonly used in machine learning: long short-term memory units [82,87]. In this setup, the 'outer loop' of learning involves tuning of the network's connection weights by a standard deep RL algorithm. Over the course of many tasks, this slowly gives rise to an 'inner loop' learning algorithm, which is implemented in the activity dynamics of the recurrent network [37,38,50,82,87]. A simple illustration involves training a recurrent neural network on a 'two-armed bandit' task. On every trial the agent must select between two actions (pull the left arm or right arm), and is rewarded based on latent parameters that govern the payoff probability for each action. During training, a series of bandit problems is presented, each with a randomly sampled pair of payoff parameters. The agent interacts with one bandit problem for a fixed number of steps and then moves on to another. The challenge for the agent, in each new bandit problem, is to balance exploration of the two alternatives with exploitation of information gained so far, seeking to maximize cumulative reward. Meta-RL learns to solve this problem, learning to explore more on a difficult bandit problem with arm reward probabilities that are more closely matched [(40%, 60%) versus (25%, 75%); compare lower panel with upper panel of Figure IA]. After training on a number of problems, the network can, even with its connection weights fixed, explore a new bandit problem and converge on the richer arm using a strategy that compares favorably with hand-crafted machine-learning algorithms (Figure IB).

Critically, the learning algorithm that arises in the network's activity dynamics is not only independent of the 'outer loop' RL algorithm that trains the network's weights; it can also work better than that outer loop algorithm, because it learns inductive biases that align with the tasks on which the system is trained. An illustration is provided in Figure IB (from Wang *et al.* [49]). The green line depicts performance (cumulative regret; lower is better) after training on a set of bandit tasks across which the arm reward probabilities are sampled independently ('independent bandits'), whereas the blue line represents training on a distribution in which the arm probabilities are linked ('correlated bandits'; specifically, they always add to one). As the figure shows, training on correlated bandits leads the algorithm to hone in on the superior arm more rapidly. This reflects an adaptive inductive bias, implemented in the network's recurrent dynamics. Figure IC illustrates a richer application of meta-RL. As described in the main text, Harlow [34] introduced the idea of 'learning to learn', using a task where monkeys chose between pairs of unfamiliar objects, eventually showing one-shot learning based on consistent structure in the task. Specifically, for each pair of objects, exactly one is consistently associated with reward, regardless of location. The monkeys eventually learned to randomly select an object on trial 1, and then use the reward outcome to perform perfectly for trials 2–6 (Figure IC, middle panel). In recent computational modeling work, Wang and colleagues [49] showed that meta-RL, employing a recurrent neural network, gives rise to the same pattern of one-shot learning (Figure IC, right panel).

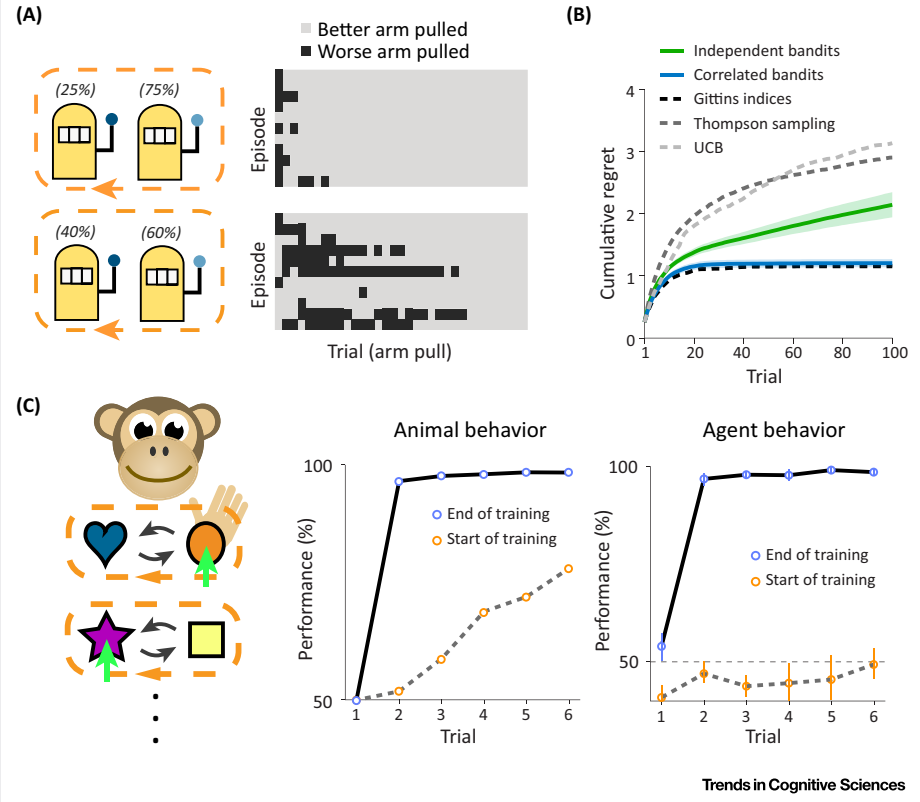

Figure I. Bandits and the Harlow Task. (A) Example behavior of meta-reinforcement learning on two-armed bandit problems at two difficulty levels. (B) Performance on correlated (blue) and independent (green) bandit problems, compared with performance of standard machine learning algorithms. (C) The Harlow task, illustration and comparison of animal to agent behavior over the course of training. Abbreviation: UCB, Upper Confidence Bound algorithm.

psychology and neuroscience, is that they undermine this argument, by showing that deep RL in fact need not be slow. This demonstration goes some distance toward defending deep RL as a potential model of human and animal learning. Beyond this general point, however, the specifics of episodic deep RL and meta-RL also point to interesting new hypotheses in psychology and neuroscience.

To begin with episodic deep RL, we have already noted the interesting connection between this and classical instance-based models of human memory, where cognition operates over stored information about specific previous observations [31,32]. Episodic RL offers a possible account for how instance-based processing might subserve reward-driven learning. Intriguingly, recent work on RL in animals and humans has increasingly emphasized the potential contribution of episodic memory, with evidence accruing that estimates of state and action value are based on retrieved memories for specific past action-outcome observations [41–44]. Episodic deep RL provides a forum for exploring how this general principle might scale to rich, high-dimensional sequential learning problems. Perhaps more fundamentally, it highlights the important role that representation learning and metric learning might play in RL based on episodic memory. Episodic deep RL thus suggests that it may be fruitful to investigate the way that fast episodic RL in humans and animals may

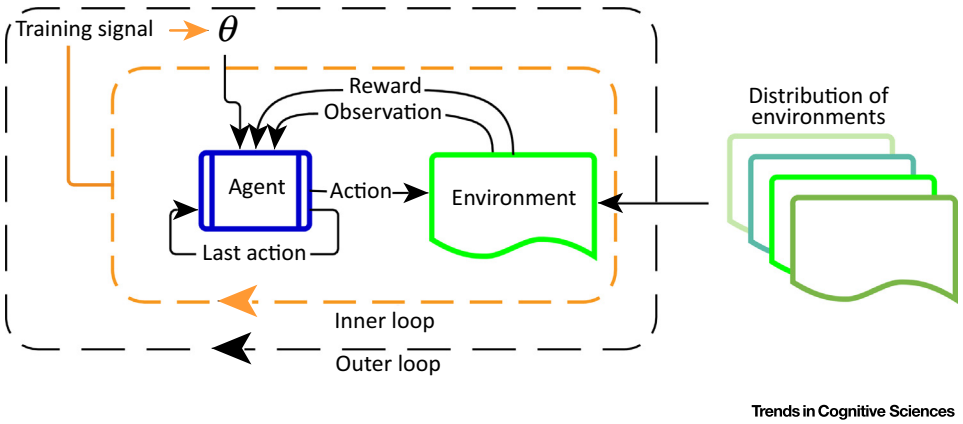

Figure 1. Schematic of Meta-reinforcement Learning, Illustrating the Inner and Outer Loops of Training. The outer loop trains the parameter weights θ, which determine the inner-loop learner ('Agent', instantiated by a recurrent neural network) that interacts with an environment for the duration of the episode. For every cycle of the outer loop, a new environment is sampled from a distribution of environments, which share some common structure.

interact with and depend upon slower learning processes. While this link between fast and slow learning has been engaged in work on memory *per se*, including work on multiple memory systems [26,45–47], its role in reward-based learning has not been deeply explored (although see [48]).

Moving to meta-RL, this framework also has interesting potential implications for psychology and neuroscience. Indeed, Wang and colleagues [49] have proposed a very direct mapping from the elements of meta-RL to neural structures and functions. Specifically, they propose that slow, dopamine-driven synaptic change may serve to tune the activity dynamics of prefrontal circuits, in such a way that the latter come to implement an independent set of learning procedures (Box 4). Through a set of computer simulations, Wang and colleagues [49] demonstrated how meta-RL, interpreted in this way, can account for a diverse range of empirical findings from the behavioral and neurophysiology literatures (Box 4).

Equally direct links connect episodic meta-RL with psychology and neuroscience. Indeed, the reinstatement mechanism involved in episodic meta-RL was directly inspired by neuroscience data indicating that episodic memory circuits can serve to reinstate patterns of activation in cerebral cortex, including areas supporting working memory (see [40]). Ritter and colleagues [39] (S. Ritter, PhD Thesis, Princeton University, 2019) show how such a function could itself be configured through RL, giving rise to a system that can strategically reinstate information about tasks encountered earlier (see also [50–52]). In addition to the initial inspiration it draws from neuroscience, this work links back to biology by providing a parsimonious explanation for recently reported interactions between episodic and model-based control in human learning ([53], S. Ritter, PhD Thesis, Princeton University, 2019). On a broader level, the work reported by Ritter and colleagues [39] provides an illustration of how meta-learning can operate over multiple memory systems [26], slowly tuning their interactions so that they collectively support rapid learning.

Fast and Slow RL: Broader Implications

In discussing both episodic RL and meta-RL, we have emphasized the role of 'slow' learning in enabling fast, sample-efficient learning. In meta-RL, as we have seen, the role of slow, weight-based learning is to establish inductive biases that can guide inference and thus support rapid

adaptation to new tasks. The role of slow, incremental learning in episodic RL can be viewed in related terms. Episodic RL inherently depends on judgments concerning resemblances between situations or states. Slow learning shapes the way that states are internally represented and thus puts in place a set of inductive biases concerning which states are most closely related.

Looking at episodic RL more closely, one might also say that there is an inductive bias built into the learning architecture. In particular, episodic RL inherently assumes a kind of smoothness principle: similar states will generally call for similar actions. Rather than being learned, this inductive bias is wired into the structure of the learning system that defines episodic RL. In current AI parlance, this is a case of 'architectural' or 'algorithmic bias', in contradistinction to 'learned bias', as seen in meta-RL.

A wide range of current AI research is focused on finding useful inductive biases to speed learning, whether this is accomplished via learning or through the direct hand-design of architectural or algorithmic biases. Indeed, the latter strategy has itself been largely responsible

Box 4. Meta-RL and Rapid Learning in the Brain

Wang and colleagues [49] proposed that meta-RL, as described in Box 3, can model learning in biological brains. They argue that recurrent networks centered on prefrontal cortex (PFC) implement the inner loop of learning, and that this inner loop algorithm is slowly sculpted by an outer loop of dopamine-driven synaptic plasticity.

We first focus on the inner loop. PFC is central to rapid learning [46,88], and neurons in PFC code for variables that underpin such learning. For example, Tsutsui *et al.* [89] recorded from primate dorsolateral PFC (dlPFC) during a foraging task in which environmental variables are continually changing. They found that individual neurons coded not only for the value of a current option, but also for the previous action taken, the previous reward received, and the interaction of previous action and previous reward (Figure I). These are key variables for implementing an effective learning policy in this task.

Wang *et al.* [49] trained meta-RL (Box 3) on the same task. They found that, over the course of training, the artificial recurrent network came to implement a fast learning policy with behavior that mimicked animals. Moreover, individual units in this trained network had coding properties remarkably similar to primate dlPFC neurons (Figure I). Wang *et al.* [49] proposed that meta-learning over the course of an animal's lifetime shapes the recurrent networks in PFC to have tuning properties that are useful for naturalistic learning tasks.

We now turn to the outer loop. Classically, midbrain dopamine neurons are thought to carry the reward prediction error (RPE) signal of temporal difference learning [90–92]. In this standard theory, dopamine drives incremental adjustments to cortico-striatal synapses, and these adjustments predispose the animal to repeat reinforced actions. This model-free learning system is often viewed as complementary to a model-based system living in mostly distinct brain areas [93–95].

However, this standard theory is complicated by the observation that dopamine carries model-based information [96–98]. For example, in a task designed to isolate model-based and model-free signals, Daw *et al.* [96] found that, surprisingly, signals in ventral striatum (thought to be the strongest fMRI correlate of dopaminergic prediction errors) contained robust model-based information (Figure I).

Wang *et al.* [49] trained meta-RL on a variant of this task [99]. In terms of behavior, they found that the trained recurrent network instantiated in its activity dynamics a robustly model-based learning procedure, despite the fact that the training procedure of meta-RL was model-free. When they examined the RPEs of meta-RLs, they found that the RPEs contained significant model-based information (Figure I). Much work has been devoted to understanding how the brain learns and uses models [46,100–102]. Meta-RL provides a single elegant explanation: that dopamine-dependent model-free learning automatically shapes recurrent networks to become a model-based learning algorithm.

In summary, meta-RL may be an important principle of brain organization. In this model, slow, incremental learning processes shape recurrent brain circuits into algorithms that exploit consistent structure in the environment to learn efficiently (for related work, see [86,103,104]).

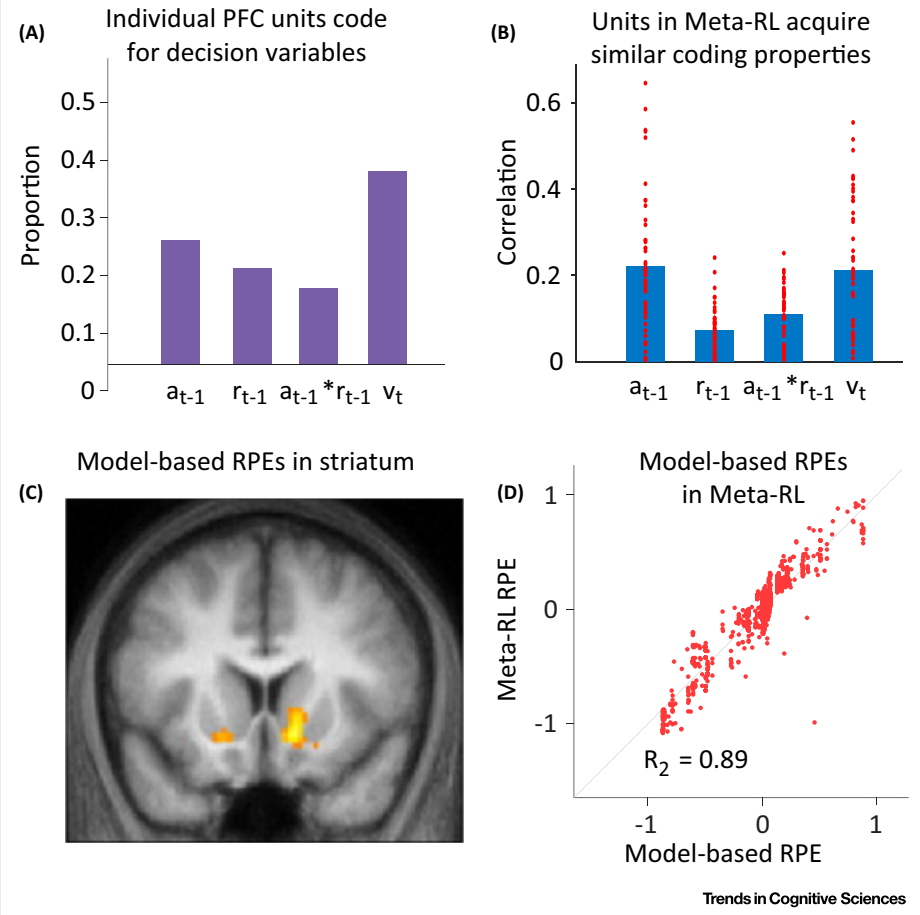

Figure I. Meta-RL in the Brain. (A,B) Proportion of units coding for task-related variables: last action, last reward, interaction of last action and last reward, and current value. (C,D) Like reward prediction errors (RPEs) measured in the brain with fMRI, RPEs in meta-reinforcement learning (RL) contain model-based information. Reproduced from [49]. Abbreviation: PFC, prefrontal cortex.

for the current resurgence of neural networks in AI. Convolutional neural networks, which provided the original impulse for this resurgence, build in a very specific architectural bias connected with translational invariance in image recognition. However, over the past few years, an increasingly large portion of AI research has been focused, either explicitly or implicitly, on the problem of inductive bias (e.g., [54–56]).

At a high level, these developments of course parallel some long-standing concerns in psychology. As we have already noted, the idea that inductive biases may be acquired through learning in fact originally derives from psychology [34] and has remained an intermittent topic of psychological research (e.g., [23]). However, meta-learning in neural networks may provide a new context to explore the mechanisms and dynamics of such learning-to-learn processes, especially in the RL setting.

Psychology, and in particular developmental psychology, has also long considered the possibility that certain inductive biases are 'built in', that is, innately specified [57]. However, the more

mechanistic notion of architectural bias, and of biases built into neural network learning algorithms, have been less widely considered (although see [58–60]). Here again, current methods for deep learning and deep RL provide a toolkit that may be helpful in further exploration, perhaps, for example, in considering the implications of growing information about brain connectivity [61–63].

It is worth noting that, whereas AI work draws a sharp distinction between inductive biases that are acquired through learning and biases that are 'wired in' by hand, in a biological context a more general, unifying perspective is available. Specifically, one can view architectural and algorithmic biases as arising through a different learning process, driven by evolution. Here, evolution is the 'slow' learning process, gradually sculpting architectural and algorithmic biases that allow faster lifetime learning. On this account, meta-learning plays out not only within a single lifespan, but also over evolutionary time. Interestingly, this view implies that evolution does not select for a truly 'general-purpose' learning algorithm, but for an algorithm which exploits the regularities in the particular environments in which brains have evolved. In this context, recent developments in AI may once again prove handy in exploring implications for neuroscience and psychology. Recent AI work has delved increasingly into methods for building agent architectures, as well as reward functions, through evolutionary algorithms, inspired by natural selection [15,64–67].

Whether it focuses on hand-engineering or evolution, AI work on architectural and algorithmic bias furnishes us with a new laboratory for investigating how evolution might sculpt the nervous system to support efficient learning. Possibilities suggested by AI research include constraints on the initial pattern of network connectivity [36]; shaping of synaptic learning rules [35,68,69]; and factors encouraging the emergence of disentangled or compositional representations [54,70,71] and internal predictive models [51].

Such work, when viewed through the lens of psychology, neuroscience, and evolutionary and developmental studies, yields a picture in which learning is operating at many time-scales at once—literally running from millennia to milliseconds—with slower time-scales yielding biases that enable faster learning at levels above, and with all of this playing out across evolutionary, developmental, and diurnal time-scales, following a trajectory that is strongly influenced by the structure of the environment (see [72]). From this perspective, evolution shapes architecture and algorithm, which embed inductive biases; these then shape lifetime learning, which itself develops further inductive biases based on experience. Within this picture, evolution appears as the 'slowest' learning process in a cascade, with each layer supporting a next, 'faster' level by serving inductive biases to that next level.

As many readers will recognize, this meta-learning-based perspective is not entirely new to cognitive science. As pointed out earlier, the idea of learning to learn has held a place in psychology for decades [34,73], and hierarchical Bayesian models of learning have long emphasized closely related ideas, with prior probabilities providing the medium for inductive biases [23,74–77]. Research on biological evolution has also touched on related ideas [66]. Notwithstanding these precedents, the recent surge in AI research investigating the role of slow and fast learning in neural networks, and in RL, provides a genuinely new set of perspectives and a new set of opportunities.

Concluding Remarks and Future Directions

The rapidly developing field of deep RL holds great interest for psychology and neuroscience, given its integrated focus on representation learning and goal-directed behavior. In the present review, we have described recently emerging forms of deep RL that overcome the apparent

Outstanding Questions

Can AI methods for sample-efficient deep RL scale to the kinds of rich task environments humans cope with? In particular, can these methods engender rich abstractions of the kind that underlie human intelligence? What kind of training environments might be necessary for this?

Does flexible, sample-efficient human learning arise from mechanisms related to those currently being explored in AI? If so, what is their neural implementation? Does something like gradient descent learning, so important for current AI techniques, occur in the brain, or is the same role played by some different mechanism?

What inductive biases are most important for learning in the environment that human learners inhabit? To what extent were these acquired through evolution and genetically or developmentally specified, and to what extent are they acquired through learning?

One thing that makes human learners so efficient is that we are active, strategic information-seekers. What are the principles that structure and motivate human exploration and how can we replicate those in AI systems?

problem of sample inefficiency, allowing deep RL to work 'fast.' These techniques not only reinforce the potential relevance of deep RL to psychology and neuroscience, countering some recent skeptical assessments; they enrich the potential connections by establishing connections to themes such as episodic memory and learning to learn. Furthermore, ideas arising from deep RL research increasingly suggest concrete and specific directions for new research in psychology and neuroscience.

As we have stressed, a key implication of recent work on sample-efficient deep RL is that where fast learning occurs, it necessarily relies on slow learning, which establishes the representations and inductive biases that enable fast learning. This computational dialectic provides a theoretical framework for studying multiple memory systems in the brain, as well as their evolutionary origins. However, human learning likely involves multiple interacting processes, beyond those discussed in this review, and we imagine that any deep RL model will need to integrate all of these in order to approach human-like learning (see Outstanding Questions). At a broader level, understanding the relationship between fast and slow in RL provides a compelling, organizing challenge for psychology and neuroscience. Indeed, this may be a key area where AI, neuroscience, and psychology can synergistically interact, as has always been the ideal in cognitive science.

Acknowledgements

The authors were funded by DeepMind.

References

1. LeCun, Y. *et al.* (2015) Deep learning. *Nature*, 521, 436
2. Goodfellow, I. *et al.* (2016) In *Deep Learning* (Vol. 1), MIT Press
3. Krizhevsky, A. *et al.* (2012) Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* 1097–1105
4. Eslami, S.M.A. *et al.* (2018) Neural scene representation and rendering. *Science*, 360, 1204–1210
5. Bahdanau, D. *et al.* (2014) Neural machine translation by jointly learning to align and translate. *arXiv*, 1409.0473
6. Van Den Oord, A. *et al.* (2016) Wavenet: a generative model for raw audio. *arXiv*, 1609.03499
7. Marblestone, A.H. *et al.* (2016) Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10, 94
8. Song, H.F. *et al.* (2017) Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 6, e21492
9. Yamins, D.L.K. and DiCarlo, J.J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356
10. Sussillo, D. *et al.* (2015) A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* 18, 1025
11. Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014) Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol.* 10, e1003915
12. Sutton, R.S. and Barto, A.G. (2018) *Reinforcement Learning: An Introduction,* MIT Press
13. Mnih, V. *et al.* (2015) Human-level control through deep reinforcement learning. *Nature*, 518, 529
14. Moravčík, M. *et al.* (2017) Deepstack: expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356, 508–513
15. Jaderberg, M. *et al.* (2018) Human-level performance in first-person multiplayer games with population-based deep reinforcement learning. *arXiv*, 1807.01281
16. Silver, D. *et al.* (2016) Mastering the game of go with deep neural networks and tree search. *Nature*, 529, 484
17. Silver, D. *et al.* (2017) Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv*, 1712.01815
18. Silver, D. *et al.* (2017) Mastering the game of go without human knowledge. *Nature*, 550, 354
19. Silver, D. *et al.* (2018) A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362, 1140–1144
20. Sutton, R.S. and Barto, A.G. (1981) Toward a modern theory of adaptive networks: expectation and prediction. *Psychol. Rev.* 88, 135
21. Schultz, W. *et al.* (1997) A neural substrate of prediction and reward. *Science*, 275, 1593–1599
22. Tsividis, P.A. *et al.* (2017) Human learning in Atari. *2017 AAAI Spring Symposium Series*,
23. Lake, B.M. *et al.* (2015) Human-level concept learning through probabilistic program induction. *Science*, 350, 1332–1338
24. Marcus, G. (2018) Deep learning: a critical appraisal. *arXiv*, 1801.00631
25. Mnih, V. *et al.* (2013) Playing atari with deep reinforcement learning. *arXiv*, 1312.5602
26. Kumaran, D. *et al.* (2016) What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends Cogn. Sci.* 20, 512–534
27. Hardt, M. *et al.* (2015) Train faster, generalize better: stability of stochastic gradient descent. *arXiv*, 1509.01240
28. Bishop, C.M. (2006) *Pattern Recognition and Machine Learning (information science and statistics),* Springer-Verlag
29. Pritzel, A. *et al.* (2017) Neural episodic control. *arXiv*, 1703.01988
30. Gershman, S.J. and Daw, N.D. (2017) Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* 68, 101–128
31. Logan, G.D. (1988) Toward an instance theory of automatization. *Psychol. Rev.* 95, 492
32. Smith, E.E. and Medin, D.L. (1981) In *Categories and Concepts* (Vol. 9), Harvard University Press

33. Schaul, T. and Schmidhuber, J. (2010) Metalearning. *Scholarpedia*, 5, 4650 revision #91489
34. Harlow, H.F. (1949) The formation of learning sets. *Psychol. Rev.* 56, 51
35. Andrychowicz, M. *et al.* (2016) Learning to learn by gradient descent by gradient descent. *Adv. Neural Inf. Process. Syst.* 3981–3989
36. Finn, C. *et al.* (2017) Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv*, 1703.03400
37. Wang, J.X. *et al.* (2016) Learning to reinforcement learn. *arXiv*, 1611.05763
38. Duan, Y. *et al.* (2016) RI^2: fast reinforcement learning via slow reinforcement learning. *arXiv*, 1611.02779
39. Ritter, S. *et al.* (2018) Been there, done that: meta-learning with episodic recall. *International Conference on Machine Learning*, 4351–4360
40. Ritter, S. *et al.* (2018) Episodic control as meta-reinforcement learning. *bioRxiv*, 360537
41. Gershman, S.J. and Daw, N.D. (2017) Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* 68, 101–128
42. Lengyel, M. and Dayan, P. (2008) Hippocampal contributions to control: the third way. *Adv. Neural Inf. Process. Syst.* 889–896
43. Bornstein, A.M. *et al.* (2017) Reminders of past choices bias decisions for reward in humans. *Nat. Commun.* 8, 15958
44. Bornstein, A.M. and Norman, K.A. (2017) Reinstated episodic context guides sampling-based decisions for reward. *Nat. Neurosci.* 20, 997
45. Tse, D. *et al.* (2007) Schemas and memory consolidation. *Science*, 316, 76–82
46. Behrens, T.E.J. *et al.* (2018) What is a cognitive map? organizing knowledge for flexible behavior. *Neuron*, 100, 76–82
47. McClelland, J.L. *et al.* (1995) Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychol. Rev.* 102, 419
48. Ballard, I.C. *et al.* (2018) Hippocampal pattern separation supports reinforcement learning. *bioRxiv*, 293332
49. Wang, J.X. *et al.* (2018) Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* 21, 860
50. Santoro, A. *et al.* (2016) Meta-learning with memory-augmented neural networks. *International Conference on Machine Learning*, 1842–1850
51. Wayne, G. *et al.* (2018) Unsupervised predictive memory in a goal-directed agent. *arXiv*, 1803.10760
52. Graves, A. *et al.* (2016) Hybrid computing using a neural network with dynamic external memory. *Nature*, 538, 471
53. Vikbladh, O. *et al.* (2017) Episodic contributions to model-based reinforcement learning. *Annual Conference on Cognitive Computational Neuroscience CCN*,
54. Battaglia, P.W. *et al.* (2018) Relational inductive biases, deep learning, and graph networks. *arXiv*, 1806.01261
55. Kulkarni, T.D. *et al.* (2016) Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation. *Adv. Neural Inf. Process. Syst.* 3675–3683
56. Vinyals, O. *et al.* (2016) Matching networks for one shot learning. *Adv. Neural Inf. Process. Syst.* 3630–3638
57. Spelke, E.S. and Kinzler, K.D. (2007) Core knowledge. *Dev. Sci.* 10, 89–96
58. Botvinick, M.M. (2007) Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 362, 1615–1626
59. Rougier, N.P. *et al.* (2005) Prefrontal cortex and flexible cognitive control: rules without symbols. *Proc. Natl. Acad. Sci. U. S. A.* 102, 7338–7343
60. Plaut, D.C. (2002) Graded modality-specific specialisation in semantics: a computational account of optic aphasia. *Cogn. Neuropsychol.* 19, 603–639
61. Sporns, O. *et al.* (2004) Organization, development and function of complex brain networks. *Trends Cogn. Sci.* 8, 418–425
62. Bullmore, E. and Sporns, O. (2012) The economy of brain network organization. *Nat. Rev. Neurosci.* 13, 336
63. Modha, D.S. and Singh, R. (2010) Network architecture of the long-distance pathways in the macaque brain. *Proc. Natl. Acad. Sci. U. S. A.* 107, 13485–13490
64. Stanley, K.O. and Miikkulainen, R. (2002) Evolving neural networks through augmenting topologies. *Evol. Comput.* 10, 99–127
65. Fernando, C.T. *et al.* (2018) Meta learning by the Baldwin effect. *arXiv*, 1806.07917
66. Hinton, G.E. and Nowlan, S.J. (1987) How learning can guide evolution. *Complex Syst.* 1, 495–502
67. Wang, J.X. *et al.* (2018) Evolving intrinsic motivations for altruistic behavior. *arXiv*, 1811.05931
68. Bengio, Y. *et al.* (1991) Learning a synaptic learning rule. In *Proceedings from IJCNN-91 Seattle International Conference on Neural Networks*, pp. 696, IEEE 2
69. Jaderberg, M. *et al.* (2016) Decoupled neural interfaces using synthetic gradients. *arXiv*, 1608.05343
70. Higgins, I. *et al.* (2016) Beta-vae: learning basic visual concepts with a constrained variational framework. *5th International Conference on Learning Representations*,
71. Higgins, I. *et al.* (2017) Darla: improving zero-shot transfer in reinforcement learning. *arXiv*, 1707.08475
72. Botvinick, M. *et al.* (2015) Reinforcement learning, efficient coding, and the statistics of natural tasks. *Curr. Opin. Behav. Sci.* 5, 71–77
73. Lake, B. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, E253 http://dx.doi.org/10.1017/S0140525X16001837
74. Griffiths, T.L. *et al.* (2008) *Bayesian Models of Cognition. Cambridge Handbooks in Psychology,* pp. 59–100, Cambridge University Press
75. Griffiths, T.L. *et al.* (2010) Probabilistic models of cognition: exploring representations and inductive biases. *Trends Cogn. Sci.* 14, 357–364
76. Tenenbaum, J.B. *et al.* (2011) How to grow a mind: statistics, structure, and abstraction. *Science*, 331, 1279–1285
77. Grant, E. *et al.* (2018) Recasting gradient-based meta-learning as hierarchical bayes. *arXiv*, 1801.08930
78. Tesauro, G. (1995) Temporal difference learning and td-gammon. *Commun. ACM*, 38, 58–68
79. Hessel, M. *et al.* (2017) Rainbow: combining improvements in deep reinforcement learning. *arXiv*, 1710.02298
80. Jaderberg, M. *et al.* (2017) Population based training of neural networks. *arXiv*, 1711.09846
81. Blundell, C. *et al.* (2016) Model-free episodic control. *arXiv*, 1606.04460
82. Hochreiter, S. *et al.* (2001) Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pp. 87–94, Springer
83. Thrun, S. and Pratt, L. (1998) Learning to learn: introduction and overview. In *Learning to Learn*, pp. 3–17, Springer
84. Baxter, J. (1998) Theoretical models of learning to learn. In *Learning to Learn*, pp. 71–94, Springer
85. Schmidhuber, J. *et al.* (1996) *Simple principles of metalearning. Technical report,* SEE
86. Schweighofer, N. and Doya, K. (2003) Meta-learning in reinforcement learning. *Neural Netw.* 16, 5–9
87. Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. *Neural Comput.* 9, 1735–1780
88. Rushworth, M.F.S. *et al.* (2011) Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70, 1054–1069
89. Tsutsui, K.-I. *et al.* (2016) A dynamic code for economic object valuation in prefrontal cortex neurons. *Nat. Commun.* 7, 12554

90. Montague, P.R. *et al.* (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* 16, 1936–1947

91. Glimcher, P.W. (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* 108, 15647–15654

92. Watabe-Uchida, M. *et al.* (2017) Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* 40, 373–394

93. Dolan, R.J. and Dayan, P. (2013) Goals and habits in the brain. *Neuron*, 80, 312–325

94. Daw, N.D. *et al.* (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* 8, 1704

95. Balleine, B.W. and Dickinson, A. (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37, 407–419

96. Daw, N.D. *et al.* (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69, 1204–1215

97. Sadacca, B.F. *et al.* (2016) Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife*, 5, e13665

98. Sharpe, M.J. *et al.* (2017) Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* 20, 735

99. Miller, K.J. *et al.* (2017) Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* 20, 1269

100. Gläscher, J. *et al.* (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66, 585–595

101. Kolling, N. *et al.* (2016) Value, search, persistence and model updating in anterior cingulate cortex. *Nat. Neurosci.* 19, 1280

102. O'Reilly, R.C. and Frank, Michael J. (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* 18, 283–328

103. Ishii, S. *et al.* (2002) Control of exploitation-exploration meta-parameter in reinforcement learning. *Neural Netw.* 15, 665–687

104. Khamassi, M. *et al.* (2013) Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. In *Progress in Brain Research* (Vol. 202), pp. 441–464, Elsevier

105. McClelland, J.L. and Rumelhart, D.E. (1989) *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises,* MIT Press