nature machine intelligence

REVIEW ARTICLE

https://doi.org/10.1038/s42256-019-0025-4

Reinforcement learning in artificial and biological systems

Emre O. Neftci[1,3] and Bruno B. Averbeck[2,3]*

There is and has been a fruitful flow of concepts and ideas between studies of learning in biological and artificial systems. Much early work that led to the development of reinforcement learning (RL) algorithms for artificial systems was inspired by learning rules first developed in biology by Bush and Mosteller, and Rescorla and Wagner. More recently, temporal-difference RL, developed for learning in artificial agents, has provided a foundational framework for interpreting the activity of dopamine neurons. In this Review, we describe state-of-the-art work on RL in biological and artificial agents. We focus on points of contact between these disciplines and identify areas where future research can benefit from information flow between these fields. Most work in biological systems has focused on simple learning problems, often embedded in dynamic environments where flexibility and ongoing learning are important, similar to real-world learning problems faced by biological systems. In contrast, most work in artificial agents has focused on learning a single complex problem in a static environment. Moving forward, work in each field will benefit from a flow of ideas that represent the strengths within each discipline.

Biological and artificial agents must achieve goals to survive and be useful. This goal-directed or hedonistic behaviour is the foundation of reinforcement learning (RL)[1], which is learning to choose actions that maximize rewards and minimize punishments or losses. Reinforcement learning is based on interactions between an agent and its environment (Fig. 1a,b). The agent must choose actions based on sensory inputs, where the sensory inputs define the states of the environment. It is the outcomes of these actions over time, either rewards or punishments, that the agent tries to optimize. This formulation is natural for behaviour in biological systems, but it has also proven highly useful for artificial agents.

Biological systems must find food, avoid harm and reproduce[2]. The environments in which they live are dynamic and key processes unfold on multiple timescales (Fig. 1c). While some of these changes can be slow and persistent (for example, seasonal), others can be sudden and ephemeral (for example, the appearance of a predator) and even fast and persistent (for example, destruction of a habitat). To deal with these changes, biological systems have to continuously adapt and learn on multiple timescales. Studies of biological systems have often focused on understanding how organisms deal with learning problems where the associations between choices and rewards are immediate, but dynamic[3–5]. These are similar to ecological problems like learning which food to eat and whether conspecifics are friendly. The values assigned to choices in these cases can be updated rapidly with experience because the credit assignment problem—the link between the choice and the outcome—is straightforward. More concretely, in two-armed bandit paradigms often used to study learning in animals, the rewards associated with choice options can be learned rapidly, and updated when they change[6,7].

On the other hand, artificial agents are constructed from mathematical models and typically trained to solve a single problem in a static environment[8,9], meaning that the reward contingencies and environmental responses are statistically fixed. In recent years, the most successful artificial systems, including neural networks, are generally trained in a data-driven fashion through statistical optimization[10]. Training on these problems takes an enormous number of trials (Fig. 1d). Due to specific requirements for optimization, the training phase is generally separated from the performance phase (Fig. 1f). The separation of training and performance prevents artificial agents from benefiting from ongoing experience or adapting to changes in the environment. As we discuss later, joining these two phases to form one single 'lifelong learner' can lead to instabilities that challenge the assumptions made in statistical learning. Researchers are now attempting to address these issues (for example, DARPA's Life Learning Machines (L2M) programme and DeepMind), using approaches like multitask reinforcement learning[11,12]. However, achieving the data efficiency and adaptability of biological agents in dynamical environments remains a major challenge.

Despite differences between work on learning in biological and artificial agents, or perhaps due to these differences, there is much room for the flow of ideas between these fields. Systems neuroscience has used many theoretical concepts from the study of RL in artificial agents to frame questions about biological systems. Theoretical RL algorithms, both model-free and model-based, are providing novel insights into reward-based learning processes in biology[13,14]. Moving from biology to theory, much of the work on learning in artificial neural networks was driven by ideas from learning in biology, including the perceptron[15] and the wake–sleep algorithm[16], which laid the foundations for efficient training of deep networks today.

A growing body of work explores the intersection of learning in artificial and biological systems. This work attempts on the one side to build an artificial brain and on the other to understand biological brains. In this Review, we focus on describing areas where the flow of ideas from the study of learning in artificial systems has led to increased understanding of learning in biological systems, and vice versa. We also point to areas where this flow of ideas may be exploited in the future, to better understand biological learning

[1]Department of Cognitive Sciences, Department of Computer Science, University of California Irvine, Irvine, CA, USA. [2]Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, MD, USA. [3]These authors contributed equally: Emre O. Neftci, Bruno B. Averbeck. *e-mail: bruno.averbeck@nih.gov

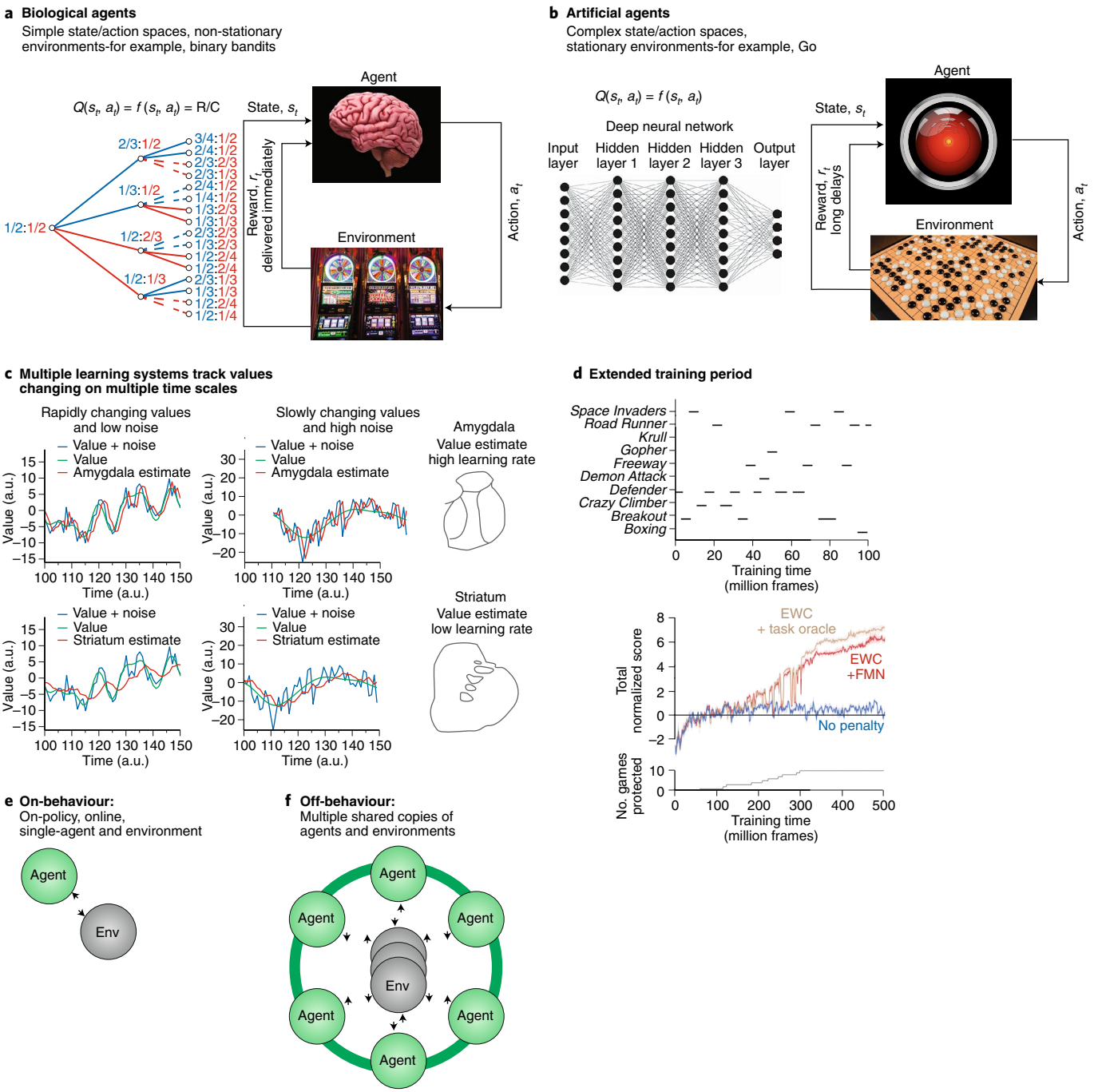

Fig. 1 | Overview of approaches to learning in biological and artificial agents. a, RL is based on the interaction between an agent and its environment. Agents choose actions, a_t, which lead to changes of the state, s_t, and rewards, r_t, which are returned by the environment. The agent's goal is to maximize rewards over a defined time horizon. Action values, $Q(s_t, a_t)$ in experiments used to study RL in biology are often simple functions of the frequencies of choices and rewards (that is, number of rewards R divided by the number of choices C for a small number of choices in bandit tasks). **b**, The same agent–environment distinction is important in artificial systems. In state-of-the-art artificial RL systems, action values are estimated by training deep networks. They are often complex functions of sensory inputs. **c**, Biological agents (for example, the brain) employ multiple learning systems that learn at different rates. The amygdala and striatum are two nuclei in the brain that can support RL learning. The amygdala (also see Fig. 3) learns at a fast rate, and therefore can track rapid changes in the environment, but at the expense of sensitivity to noise. The striatum, on the other hand, learns more slowly. While it cannot track rapid changes in environmental values, it is more robust to noise. **d**, Artificial agents are often trained on complex, statistically stationary problems. The number of training trials is huge, and therefore these systems cannot adapt rapidly to changes in the environment. Artificial agents are often trained on a single task and fail to learn in sequential multitask settings. Hierarchical RL, structural plasticity and consolidation can enable artificial agents to learn on multiple timescales. **e**, Biological agents interact with the environment in an 'on-behaviour' fashion—that is, learning is online and there is a single copy of the environment. **f**, While many RL approaches for artificial agents follow these principles, the most recent and successful strategies include a form of agent parallelism, where the agents learn on copies of the environment to stabilize learning (see, for example, A3C and IMPALA). Experience replays inspired by the hippocampus or more complementary learning systems can provide the necessary properties for on-behaviour agents, and thus form a point of contact between artificial and biological RL. Credit: Sebastian Kaulitzki/Alamy Stock Photo (brain image); Moritz Wolf/imageBROKER/Alamy Stock Photo (slot machine image).

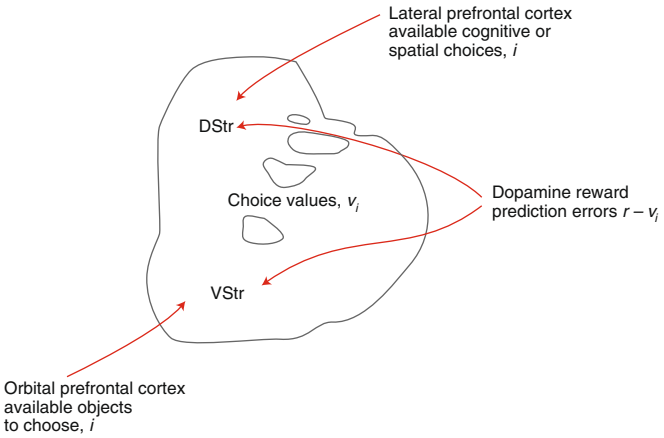

Fig. 2 | Anatomy of a model of reinforcement learning shown on a schematic representation of the rhesus monkey striatum. The model is focused on dopamine and its effects in the striatum. DStr: dorsal striatum. VStr: ventral striatum. Red lines indicate anatomical inputs from the indicated neural population to the striatum. Cortical inputs are excitatory. The dopamine input arises in the midbrain dopamine neurons.

systems and to build artificial agents capable of solving increasingly complex real-world problems. Finally, we consider how these bridges underlie recent advances in engineering brain-inspired, neuromorphic technologies.

The paper is organized by first considering RL in biological systems. We start by discussing computationally simple model-free learning problems, where much is known about both the neural circuitry and behaviour, and ideas from learning in artificial agents have had a deep influence. We then move on to more complex, model-based RL, where ideas from learning in artificial agents have provided tools for understanding behaviour, and work on neural systems is just beginning. The second half of the paper is focused on learning in artificial systems. We start with an overview of the recent successes in training artificial RL systems to solve complex problems that have relied on developments in deep neural networks, which were inspired by networks in biological systems. We then discuss hierarchical RL, a framework developed for learning in artificial agents. It is likely to prove useful in the future for understanding biological systems, as there is little known about the neural circuitry that underlies hierarchical learning in biological agents. Finally, we consider neuromorphic engineering, an area of research that draws explicitly from biology to solve real-world engineering problems.

Biological systems underlying RL

The theoretical constructs of model-free and model-based reinforcement learning were developed to solve learning problems in artificial systems. They have, however, been used to understand learning problems in biological systems at both the behavioural and neural levels. The neural basis of RL in mammalian systems, particularly model-free RL, is arguably one of the best understood in systems neuroscience[17–21]. This is due to the success of temporal-difference RL[18] and Rescorla–Wagner theories for predicting the activity of dopamine neurons, and the effects of activating dopamine neurons on behaviour. Theories of model-free RL have emphasized the role of frontal-striatal systems[20,21], which are the anatomically defined networks connecting prefrontal cortex to the striatum[22], and dopamine-driven plasticity in these circuits (Fig. 2). According to one model, cortex represents the set of available choices[23]. The strength of cortical synapses on striatal cells encodes information about the values of each of the choices[21]. Stronger synapses drive increased activity in striatal cells. Therefore, the activity of striatal cells represents the values of the options represented by cortex[24,25]. The striatal activity drives choice activity, either via downstream circuitry through the basal ganglia and return loops through the thalamus to the cortex, or via descending projections to brain-stem motor output areas. After making a choice and experiencing an outcome, dopamine encodes a reward prediction error, RPE$=r - v_i$. The RPE is the difference between the expected value of the chosen option, v_i, encoded by the striatum, and the experienced outcome, r. If the RPE is positive, the outcome was better than expected, and there is a phasic increase in dopamine. If the RPE is negative, the outcome was worse than expected, and there is a phasic decrease in dopamine. This change in dopamine concentration drives plasticity on the frontal-striatal synapses representing the chosen option. Increases in dopamine drive increases in synaptic strength and decreases in dopamine drive decreases in synaptic strength (ignoring for simplification direct and indirect pathways). The next time these choice options are experienced, the activity of the striatal neurons will reflect this updated synaptic efficacy, firing more for options that had a positive RPE in the previous trial, and less for options that had a negative RPE. This process in its simplest form is captured by the Rescorla–Wagner equation, which is a stateless RL update model[17,26].

$$v_i(k+1) = v_i(k) + \alpha(r(k) - v_i(k))$$

This equation summarizes the interaction of neural activity in three brain areas—cortex, which represents the options, i; the striatum, which represents their values v_i; and mid-brain dopamine neurons, which code RPEs. The equation further describes at a formal level the process of changing value representations during learning that underlie behaviour, where the size of the update is controlled by a learning rate parameter, α. (Note that the original Rescorla–Wagner equation was developed in the context of Pavlovian cue conditioning and not choices.)

Temporal-difference (TD) learning, first developed for artificial systems[27], provides an extension of the Rescorla–Wagner model, to cases where action values depend on states. The state is defined by the information relevant to choosing an option and can be, for example, time. The TD update rule for actions, i, is given by

$$v_i(s_t) \leftarrow v_i(s_t) + \alpha(r(t) - v_i(s_t) + \gamma v_i(s_{t+1}))$$

In this case, we have used the assignment operator $\leftarrow$ to indicate the update after an event. The variable s_t is the state at time t, and γ is a discount parameter that discounts the value of future states. The TD RPE is given by

$$\delta(t) = r(t) - v_i(s_t) + \gamma v_i(s_{t+1})$$

This general theory has been highly successful, and it predicts much behavioural and neural data. For example, a substantial body of work has shown that dopamine neurons code TD RPEs under diverse conditions[28], and activating dopamine neurons is equivalent to experiencing an RPE, with respect to learning[29–31].

However, this model leaves many details unspecified. For example, it is now clear that a larger set of interconnected areas underlies RL (Fig. 3). These networks are organized around a set of overlapping but segregated cortical-basal ganglia-thalamo-cortical systems. Broadly speaking, there is one system interconnected with the dorsal striatum that mediates learning about rewarding spatial-cognitive processes—for example, spatially directed eye-movements[32–34]—and another system interconnected with the ventral striatum that mediates learning about rewarding stimuli, particularly visual stimuli in the primate[35,36]. The ventral system also has strong inputs from the amygdala, which plays an important role in learning about the values (positive and negative) of stimuli in the environment[26,37].

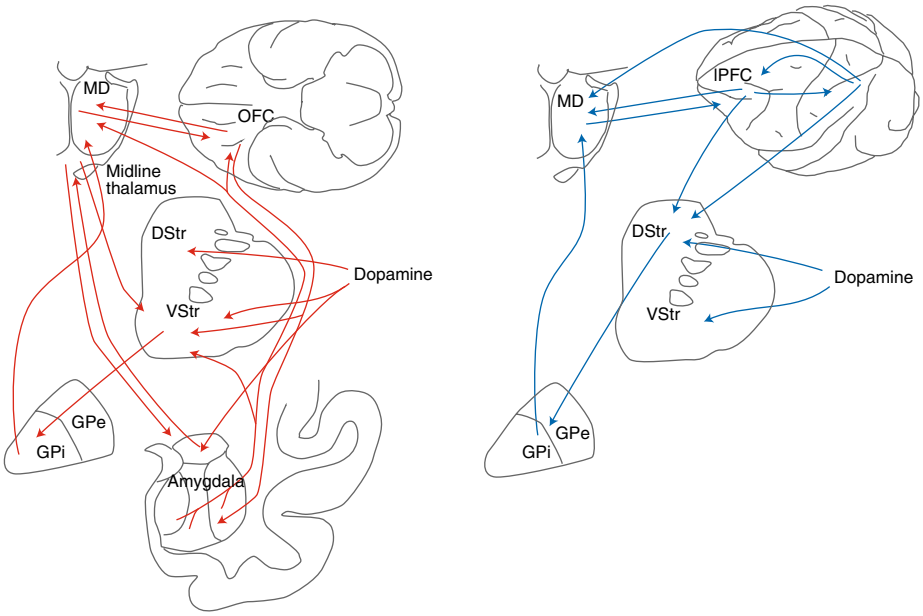

Fig. 3 | Expanded conception of neural circuitry underlying reinforcement learning. Example systems are taken from a rhesus monkey. The lines indicate anatomical connections between the indicated regions. The network on the left, in red, is specialized for associating values to objects defined by sensory information. This network interconnects mostly ventral structures, including orbito-frontal cortex and the ventral striatum. The network on the right, in blue, is specialized for associating values to cognitive and spatial processes. This network interconects dorsal systems, including dorsal-lateral prefrontal cortex and the dorsal striatum. These systems have a parallel organization, such that the circuits from the cortex, through the basal ganglia (striatum and GPi) and thalamus and back to cortex are organized similarly. The amygdala, however, only participates in the sensory network shown on the left. OFC: orbital prefrontal cortex. lPFC: lateral prefrontal cortex. MD: medial-dorsal thalamus. GPe/GPi: globus pallidus external and internal segments.

There does not appear to be a corresponding structure for the dorsal circuit. The hippocampus, for example, which is not included here, projects to the ventral striatum, not the dorsal striatum[37]. It is also important to point out that while we present the dorsal and ventral systems as separate circuits, there is a gradient of circuitry interconnecting all parts of the prefrontal cortex to corresponding areas in the striatum[35,38]. The dorsal and ventral circuits represent two poles in the system. Furthermore, the role of the dorsal system in eye movements and the role of the ventral system in vision follows to some extent from the tasks that have been used to study them.

Multiple timescales of learning

Biological agents must solve learning problems on multiple timescales and the neural systems underlying RL in biological agents reflect this need. In the ventral system (Fig. 3, red lines), there is an interacting and parallel organization of the amygdala and striatum. Work that has examined the relative contribution of these two systems to RL has suggested that plasticity within the amygdala operates on fast timescales, through an activity-dependent mechanism, whereas plasticity in the striatum operates on slower timescales, through a dopamine-dependent mechanism[7,26]. Having parallel neural systems that learn at different rates allows biological agents to learn efficiently and track changing values in environments characterized by non-stationarities on different timescales[39]. The slower striatal/dopamine-dependent system learns more effectively in noisy environments (Fig. 1c), when the values of choices evolve on slower timescales. The amygdala/activity-dependent system, on the other hand, learns more effectively when environments and the underlying values of choices evolve more quickly. However, the amygdala system is more susceptible to noise. The amygdala, due to its rapid activity-dependent plasticity mechanisms, erroneously tracks noisy fluctuations in values, which can lead to inaccurate value estimates if noise is large relative to signal. The striatum, because it updates values slowly, tends to integrate out this noise. Because these two systems both track values, a downstream system must mediate between them, combining the value estimates from each system, according to an ongoing reliability estimate. This mediation process is known as mixture-of-experts in machine learning, where the concept was first developed[40]. If one of the systems is providing more accurate value estimates, its contribution to behaviour should be up-weighted, relative to the other. It is currently not clear where this downstream system is, although it may be in cortical motor structures, and therefore effector specific. Overall, however, this organization reflects a general principle underlying the biological solutions to RL problems. Specifically, the brain uses multiple interconnected systems to solve the RL problem (Fig. 3). It decomposes the problem into sub-problems and uses multiple parallel but interacting systems to solve learning problems flexibly.

In computational neuroscience, studies of synaptic plasticity have shown that the timescales of learning are directly related to the dynamical processes of neurons and synapses[41]. Specifically, these studies found that the spike-timing-dependent plasticity learning window, which determines the magnitude of weight updates (that is, synaptic plasticity), is a direct reflection of the post-synaptic potential, typically in the 1 ms to 100 ms range depending on the neuron and synapse type. More generally, the theory implies that the timescales of the plasticity processes match the timescales of neural activity. Furthermore, theoretical and modelling studies show that having both slow and fast timescales achieves extended memory capacity[42,43], improves performance[44] and speeds up learning[45]. Therefore, even at the level of single neurons, learning with multiple timescales is advantageous. The multiplicity of timescales is also a central feature of certain artificial recurrent neural networks, such as those composed of long short-term memory (LSTM) units[46]. LSTMs were originally designed to improve the temporal horizon of learning by introducing a memory element whose decay time constant is dynamic and data-dependent, inspired by working memory processes studied in biology. Interestingly, models using

working memory or multiple clock rates have been shown to reproduce some of the LSTM's computational capabilities[44,47].

The continuous operation of neural dynamics in the brain entails that, in contrast to conventional machine learning and RL in artificial agents, learning in the brain is an ongoing process. In continuous learning, parameters are updated (that is, plasticity) sequentially, following each data sample[48] (that is, 'online' in the machine learning sense). In contrast, batch updates commonly used in machine learning involve processing many events or 'trials' before connection weights are updated. This requires storage of these trials. Statistical learning theory shows that it is advantageous in gradient-based learning to operate in the sequential fashion[49], both in terms of memory and computational complexity, provided data are independent and identically distributed (iid). When the iid case is violated, correlations in the data sampling can lead to catastrophic interference[50,51]—that is, old knowledge tends to be overwritten by new knowledge. This is particularly problematic in RL because data are inherently correlated and the agent modifies the data-sampling process through the choice of actions. Two main avenues exist to combat this problem: complementary learning and replay mechanisms. With replay mechanisms, older experiences are presented to the network again[51–54]. With complementary learning, synapse-specific learning rates change according to past task relevance[12,55,56]. In this case, a network mechanism estimates the importance of neurons and synapses in a task and selectively stiffens their parameters. Both mechanisms were inspired by the brain.

Learning to learn and model-based RL

In addition to the model-free learning systems discussed above, mammalian systems can learn using more sophisticated model-based inference strategies. As a side note, the term model-based originally referred specifically to RL learning problems in which state transition functions were known, which allows for Bellman updates[57]. However, the term has come to more generally mean any learning process that relies on knowledge of the statistics of the environment, and therefore uses a statistical—usually Bayesian—model. Work in this area has borrowed extensively from concepts first developed to solve learning problems in artificial agents. There is substantial behavioural evidence for model-based inference strategies[58–61], but much less is currently known about the neural circuitry, relative to model-free learning[59,62]. The original theory of model-based RL for biological agents placed model-based learning in prefrontal cortex[57]. This was consistent with general ideas about cognitive planning processes being driven by the prefrontal cortex[63]. However, most subsequent work, and the original rat experiments on habit versus goal-directed systems that inspired the theory[64], does not support this distinction. Several studies have shown that both model-based and model-free learning rely on striatal-dependent processes[62,65], although some studies have suggested that prefrontal cortex underlies aspects of model-based learning[60]. Therefore, it is clear that biological systems can use model-based approaches to learn, but the neural systems that underlie this form of learning are not currently understood.

Behavioural evidence for model-based learning comes in at least three forms. First, mammals can learn to learn[66]. This means that the rate of learning on a new problem that is drawn from a class of problems with which one has experience improves as one is exposed to more examples from the class. Thus, the statistics of the underlying inference process, or the model that is generating the data, is learned over time. For example, in reversal learning experiments, animals are given a choice between two options, which can be two objects whose locations are randomized[62,67,68]. Choice of one of the options leads to a reward, and choice of the other option leads to no reward. Once an animal has learned to choose the better option, the choice–outcome mapping is switched, such that the previously rewarded option is no longer rewarded, and the previously unrewarded option is rewarded. (In probabilistic versions of this problem, the choices differ in the frequency with which they are rewarded when chosen, and these frequencies switch at reversal.) When the animals are exposed to a series of these reversals, the rate at which they switch preferences improves with experience. Thus, it may take five or ten trials to switch preferences the first time the contingencies are reversed. However, with sufficient experience, the animals may reverse preferences in just one or two trials. This process can be captured by a model that assumes a Bayesian prior over the probability of reversals occurring in the world[68]. The prior starts out low, since the animals have mostly been exposed to stable stimulus-outcome mappings that do not reverse. Because the prior is low, the animals require substantial evidence before they infer that a reversal has taken place. When they fail to receive a reward for a previously rewarded choice, they believe it is noise in the reward delivery process, and not an actual reversal in choice–outcome mappings. However, with experience on the task, the prior on reversals increases, and the animals require less evidence before inferring that a reversal has occurred, and therefore they reverse their choice preferences more rapidly.

In artificial agents, learning to learn has been put forward as a principled approach to transfer learning, as the ability to generalize across a class of related tasks implies that information used to solve one task has been transferred to another. This idea is fundamental to the recent meta-reinforcement learning approach[69], where synaptic plasticity driven by dopamine sets up activity-based learning in the prefrontal cortex. Interestingly, successfully transferring knowledge among a class of related problems is equivalent to generalization in the statistical machine learning sense and implies a principled solution to the catastrophic forgetting problem discussed above.

Second, and related to the first form of model-based learning, animals can use probabilistic inference, or latent state inference, to solve learning problems, when they have had adequate experience with the statistics of the problem[70,71]. With sufficient experience, animals can learn that a particular statistical model is optimal for solving an experimental problem. These models can then solve learning problems more effectively than model-free learning approaches. Probabilistic inference is guaranteed to be optimal, if the mammalian system is capable of learning the correct model[72]. In stochastic reversal learning, after the animal has learned that reversals occur, detecting a reversal statistically can be done efficiently using Bayesian inference. This is state inference, since the reward environment is in one of two states (that is, either choice one or choice two is more frequently rewarded). This process can be faster and more efficient than carrying out model-free value updates. To solve this problem with model-free value updates, the animal would have to update the value of the chosen option, using feedback, on each trial. In addition to the efficiency of Bayesian state inference, it has also been shown that animals can learn priors over reversal points, in tasks where reversals tend to happen at predictable points in time[58]. This is more sophisticated than the prior discussed above, which is a prior on the occurrence of reversals. Priors on the timing of reversals reflect knowledge that reversals tend to occur at particular points in time, and therefore implicitly assume that they occur. These priors play an important role when stochastic choice–outcome mappings make inference difficult. For example, if the optimal choice in a two-armed bandit task delivers rewards 60% of the time and the sub-optimal choice delivers rewards 40% of the time, reversals in the choice outcome mapping will be hard to detect based upon the received rewards and the priors can improve performance. It is not always straightforward, however, to dissociate fast model-free learning from model-based learning, and therefore careful task design and model fitting is required to demonstrate model-based learning in biological systems. Much of the work on these inference processes has suggested that they occur in cortex[70,71,73,74]. This raises the question of whether these processes require plasticity, or

whether they rely on faster computational mechanisms, like attractor dynamics. It is possible, for example, that the inference process drives activity in cortical networks into an attractor basin, similar to the mechanism that may underlie working memory[75,76].

A third and final form of faster learning is model-based, Bellman RL, which is known more accurately as dynamic programming. In this form of model-based learning, one has knowledge of the statistics of the environment[9,77]. These statistics include the state action reward function, $r(s_t, a)$, the state value function, $u_t(s_t)$, and the state-transition function, $p(j|s_t, a)$. When these functions are known, one can use Bellman's equation to arrive at rapid, but computationally demanding, solutions to problems.

$$u_t(s_t) = \max_{a \in A_{s_t}} \left\{ r(s_t, a) + \gamma \sum_{j \in S} p(j|s_t, a) u_{t+1}(j) \right\}$$

This equation shows that the value of the current state, s_t, is equal to the maximum over the possible actions A_{s_t}, of the immediate reward for action a, $r(s_t, a)$, plus the discounted maximum expected reward going forward from that state, which is an expected value over future states, j, from the set of reachable states S. In computational work, and correspondingly in biological systems, one rarely has access to all of the information necessary to solve Bellman's equation, and therefore model-free approaches are often necessary to learn state or action value functions. However, biological systems can learn state-transition functions, and these state-transition functions can make learning more efficient, under some conditions[59,60,78,79]. It is the case that model-free and model-based learning approaches will converge to the same result, however. Therefore, these solutions have been compared on problems that have intrinsic non-stationarities. The faster learning of the model-based system can then be revealed by its increased ability to track the non-stationarities.

Artificial connectionist RL agents

The recent successes of machine learning and RL techniques in artificial systems are spurring widespread interest in solving practical problems with artificial intelligence, such as natural language processing, speech generation, image recognition, autonomous driving and game playing. Much of this renewed interest is due to breakthroughs in deep neural networks and advances in the technologies that support them. In fact, in artificial agents, task-relevant features of the environment (states) must be inferred from high-dimensional sensory data—for example, pixel intensity values from a camera. Human observers can immediately identify the objects and their relative locations in visual data, and assign meaning to these objects. Solving these problems in artificial systems and achieving state-of-the-art performance requires specialized structures and massive amounts of training data[53]. Early pattern recognition models used hand-created features or linear models to extract states from high-dimensional sensory stimuli[80]. These methods required domain-specific knowledge and learning was limited to the selected domain. Thanks to large datasets, and improved computing technologies, deep learning was surprisingly successful in mapping high-dimensional sensory stimuli to task relevant output, or in the case of RL, in mapping from sensor data to chosen action values. Because these neural networks are general-purpose function approximators, they require few domain-specific assumptions to learn task-relevant representations of the environment[81]. Early implementations were successful at solving complex tasks, such as backgammon[82] and autonomous vehicle control[83]. With improved hardware and algorithms that prevent learning instabilities, these early approaches matured into algorithms that can now match or exceed human capabilities in a wide variety of domains, such as in game playing and motor control[53,84–86].

Can the success of deep networks and deep RL be leveraged to better understand biological agents? Interestingly, the mathematical framework of artificial recurrent neural networks can adequately describe the discrete-time approximations of simple models of biological neural networks (for example, leaky integrate-and-fire neurons). Indeed, biological neural networks are recurrent (that is, they are stateful and have recurrent connections), binary (that is, they communicate via action potentials) and operate in continuous-time (a neuron can emit an action potential at any point in time)[87], and such properties are commonly studied in artificial neural networks. One of the most constraining differences between biological and mainstream artificial learning systems is architectural: internal states such as neurotransmitter concentrations, synaptic states and membrane potentials are local. Broadly speaking, locality is characterized by the set of variables available to the processing elements (for example, the neuron and the synapse). Many critical computations in machine learning require information that is non-local—for example, to solve the credit-assignment problem. Making non-local information available to the neural processes requires dedicated channels that communicate this information[88]. The dopamine pathway is one such example. The information provided by the dopamine system is, however, only evaluative. Thus, an important challenge in bridging neuroscience and machine learning is to understand how plasticity processes can utilize this evaluative feedback efficiently for learning. Interestingly, an increasing body of work demonstrates that approximate forms of gradient backpropagation compatible with biological neural networks naturally incorporate such feedback, and models trained with them achieve near state-of-the-art results on classical classification benchmarks[89–91]. Synaptic plasticity rules can be derived from gradient descent that lead to 'three-factor' rules, consistent with an error-modulated Hebbian learning (Fig. 4). Furthermore, the normative derivation of the learning reveals plasticity dynamics that are matched to the neural and synaptic time constants discussed earlier (roughly 1 ms to 100 ms), such that spatiotemporal patterns of spikes can be efficiently learned.

Although these results have not yet been extended to deep RL, this demonstrated equivalence between biological and artificial neural networks[88] is suggestive of multiple, direct points of contact between machine learning algorithms and neuroscience. From this contact, two key challenges emerge. First, deep neural networks learning in real time require impractical amounts of experience to reach human-level classification accuracies (Fig. 1d), even on simple classical vision benchmarks and in control (the latter easily requiring millions of samples—for example, in game playing[53] or grasping with a robotic control[92]). Rapid learning is important because, behaviourally, an agent must adapt its internal representation at least as fast as the timescale of changes in the environment. The second challenge that emerges is that biological agents learn 'on-behaviour', which implies non-iid sampling of the environment or dataset, as discussed earlier (see section 'Multiple timescales of learning').

The slowness of deep learning can be partly attributed to the gradient descent-based nature of the training algorithms, which require many updates in small increments to stably approach a local minimum. Consequently, there is a fundamental trade-off in speed of learning between the amount of domain-specific knowledge that is built into a model and the number of trials necessary to learn the underlying problem. This is the same reason that model-based RL is more data efficient than model-free RL. The more domain knowledge that is built into a system, the more data efficient it will be. By design, deep RL is initialized with little or no domain-specific knowledge, and so gradient descent cannot significantly improve a deep RL model based on a single experience. In contrast, model-based RL, which incorporates problem-specific knowledge and other related methods that do not rely on gradient descent, such as tabular approaches and episodic control[1,93,94], are generally data

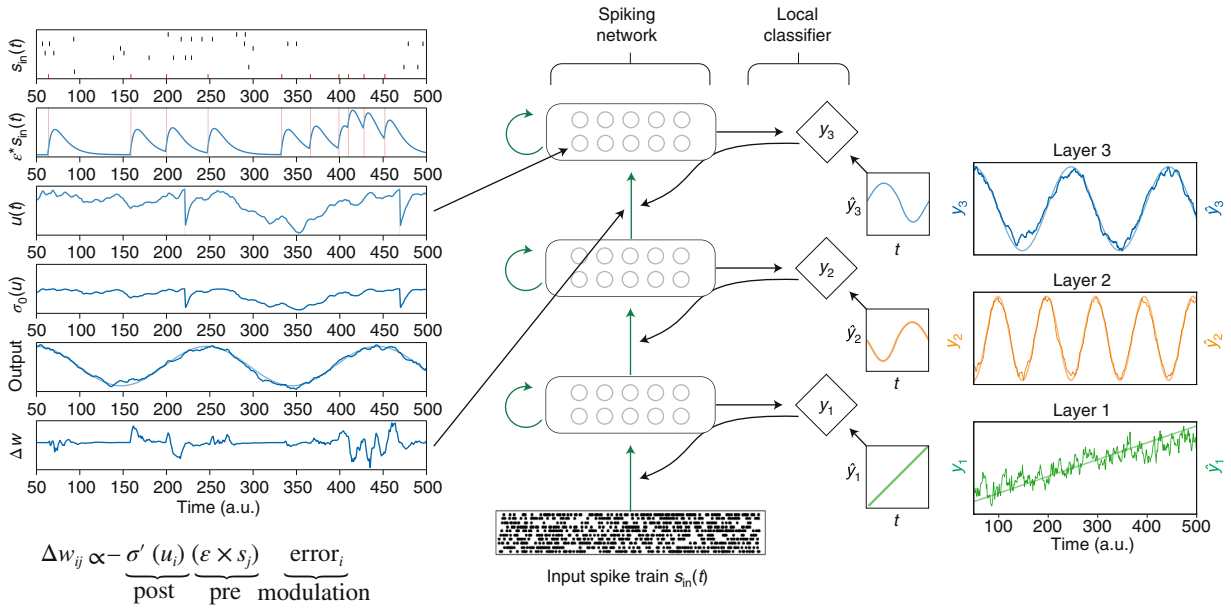

Fig. 4 | A model of spike-based deep, continuous local learning (DCLL). In a feedforward network, each layer of spiking neurons feeds additionally into a local classifier (diamond shaped) through fixed, random connections. The local classifier is trained to produce auxiliary targets $\hat{y}$. In this learning scheme, errors in the local classifiers are propagated through the random connections to train weights coming in to the spiking layer, but no further (curvy, dashed line), thus the learning is local to the layer. The synaptic plasticity rule here is derived from gradient descent. The resulting rule is composed of a pre-synaptic term (the pre-synaptic trace), a post-synaptic term (the derivative of the neural activation probability) and an error, and is consistent with a modulated Hebbian scheme. The left panel shows snapshots of the neural states during learning in the top layer, where $u(t)$ is the membrane potential and ε is the postsynaptic potential response. In this example, the network is trained to produce three time-varying auxiliary targets (targets,$\hat{y}$; predictions, y). This learning architecture and dynamics achieves near state-of-the-art classification on spatiotemporal patterns. Adapted from ref. [89] (preprint).

efficient. The complementary features of model-free and model-based RL suggest that successful artificial agents are likely to include both components (possibly in a hierarchical fashion). Meta-learning techniques that 'pre-train' a neural network on a class of related tasks[95] (in the spirit of Bayesian priors discussed earlier), hierarchy (modularity) and efficient model learning[96,97] are poised to play key roles in solving the data-efficiency problem.

The separate components of the RL problem are challenging in different ways, and therefore biological agents have evolved multiple systems optimized to the separate problems. For example, flexibly updating state values in changing environments requires solving problems that differ from state inference (for example, identification of objects and their locations)[98]. Object identification is challenging because the mapping between retinal output (in biological systems) or pixel values (in cameras) and object categories is highly nonlinear[99]. Flexibly updating state values is challenging because the information necessary to update values has to be remembered or stored over time[100], and the reward outcomes have to be attributed to the appropriate preceding decisions, to solve the credit assignment problem[101]. Bayesian inference models are also complex. For example, they often require nonlinear interactions of variables across hierarchical levels[102]. Therefore, learning based on these models requires computations of high complexity. In addition to the separation of the RL problem into state inference, value updating and action selection problems, biological agents also use multiple neural systems that are optimized to learn under different conditions. This division of labour leads to efficient solutions in biological wetware.

Hierarchical RL

Although most learning problems that have been studied in biological systems have been simple, much of behaviour, particularly human behaviour, is complex and hierarchical[103–109]. For example, when we drive to the grocery store, we do not think of every specific muscle activation necessary to first walk to the table, then pick-up our car keys, go to the door and so on. Most of the low-level behaviour is automated and these lower-level components are then sequenced at a high level. Several of these ideas have been incorporated into recent hierarchical RL algorithms. Hierarchical reinforcement learning (HRL) strives to group sequences of related low-level actions into hierarchically organized sub-goals. When tasks are abstracted into sub-goals, they can be learned more efficiently[110,111]. This is another example of building knowledge into the learning algorithm to improve data efficiency and, correspondingly, learning rates. To fully exploit the abstraction, it is customary for each level of the hierarchy to operate independently of the other and the reward. The question is then shifted to defining meaningful sub-goals that each layer of the hierarchy should solve. Because meaningful sub-goals simplify the credit-assignment problem during learning, understanding how sub-goals are learned can provide insights into biological systems. Currently, work in artificial systems leads work in biological systems in this important area. Specifically, several models have been developed to train artificial systems using hierarchical approaches. But there is little work understanding how biological systems solve these problems. The models that have been developed to solve learning problems in artificial systems are likely to prove useful for understanding these problems in biological systems.

One of the first HRL algorithms, known as the options framework, was due to Sutton et al.[111]. An option is a hand-engineered building block of behaviour and therefore incorporates substantial domain specific knowledge. For example, in a robot, an option named 'recharge' might consist of a policy for searching for a charging station, navigating towards it and initiating a docking sequence. The higher-level abstraction hides low-level actions, such as detailed navigation, and offers succinct building blocks for achieving goal-directed behaviour. In a related algorithm known as feudal RL, the system consists

of managers that learn to set tasks for their sub-managers, and sub-managers oversee those tasks while learning how to achieve them. This type of learning is interesting in distributed scenarios, as sub-managers simply need to maximize rewards in their local context, using information at their assigned level of granularity[110]. Whereas the hierarchy was fixed in the options framework, other more recent models, including feudal networks and option critics, have focused on learning this hierarchical structure. These models aim to automatically learn hierarchical task structure such as sub-policies or sub-goals[87,112]. A key challenge in HRL is to set the intermediate targets or sub-goals. One option is to use an intrinsic motivator, such as curiosity, which seeks novelty about the environment[86], or auxiliary tasks such as predicting visual features[113], to promote exploration and predict future states of the environment[114]. Another related approach is to provide a supervisory signal, such as in imitation learning[115,116] where an expert instruction is used when available.

While the study of the neural systems that underlie HRL in biological agents is just beginning, work has suggested that the same frontal networks that underlie model-free RL are relevant[107,117]. These studies have, for example, suggested that prefrontal cortex, particularly dorsal-lateral prefrontal cortex, contains a gradient, such that caudal areas implement low-level aspects of behaviours, and rostral areas implement abstracted aspects of behaviours, further up the hierarchy[103,117,118]. A major outstanding question about hierarchical control in biological systems, as is the case with artificial agents, is how they learn to group behaviours into sub-goals. Little is known about this process beyond behavioural descriptions[119]. However, the recent advances in artificial systems provide a conceptual framework for studying these problems more effectively in biological systems. For example, the auxiliary targets in deep, continuous local learning (DCLL) (Fig. 4) provide one such framework that can incorporate intermediate targets and goals.

Neuromorphic approaches

Although machine learning and neural networks share a common history with brain science, much of the recent development has strayed from these roots. A key reason for this branching is that the computers we use are different than the brain in many aspects. When some of the constraints imposed by the modern von Neumann computer architectures are relaxed, inference and learning performance can improve. For example, introducing continuous-time and online, sample-by-sample parameter updates, similar to synaptic plasticity in the brain[90], requires fewer basic operations compared to batch learning. However, on mainstream computers, neural network computations are generally carried out in batch fashion to exploit hardware parallelism, and parameter updates incur memory overhead, making near-continuous time updates suboptimal. Another example is capsule networks[120], which are best implemented on massively parallel (brain-like) hardware. These observations suggest that methods that are computationally prohibitive on conventional computers are tractable and sometimes even advantageous in massively parallel computing systems like the brain.

Neuromorphic engineering strives to bridge device physics and behaviour by taking inspiration from the brain's building blocks, such as spiking neural networks. The recent development of neuromorphic hardware and accelerators with on-chip adaptive capabilities[121–123] offers a platform for designing and evaluating brain-inspired processing and learning algorithms. Example systems were demonstrated as programmable, general-purpose sensorimotor processors[124] and reinforcement learning[125].

This hardware strives to emulate in digital or analogue technologies the dynamical and architectural properties of the brain. They consist of a large number of biologically plausible model neurons and are often equipped with synaptic plasticity to support online learning. These learning dynamics are compatible with modelling efforts in computational neuroscience, such as the three-factor learning rule sketched in Fig. 4.

While higher levels of implementation are possible to study and even implement reinforcement learning[126], these do not directly address how the realities of the physical machine, such as device-to-device variability, noise and non-locality, shape animals' inference and learning strategies. The fact that neuromorphic computing is closely dictated by its physical substrate raises computational challenges that are typically overlooked when modelling with conventional digital hardware. These challenges arise from the engineering and communication challenges of co-locating processing and memory, the energetic and hardware cost of such memory which leads to parameter and state quantization, and the unreliability of the substrate in the case of emerging devices or analogue technologies. While the spiking nature of neurons has a minor performance impact provided the credit assignment problem is addressed (Fig. 4), the quantization of the synaptic weight parameters below 8 bits of precision during learning starts to impact classification performance[127,128] and remains an open challenge. These challenges are also present in the brain, and so computational modelling at the interface of artificial and biological agents plays a key role in addressing these issues.

Neuromorphic vision sensors that capture the features of biological retinas[129] are already changing the landscape of computer vision in industry and academia. While current neuromorphic devices as general-purpose or RL processors are still in research phases, the discovery of new memory devices and the looming end of Moore's law is calling for such alternative computing strategies. Looking forward, with such hardware systems, the bridges between artificial and biological learning can directly translate into smart, adaptive technologies that can benefit medical treatments, transportation and embedded computing.

Conclusions

Artificial agents can be developed to carry out many tasks currently carried out by people. Self-driving cars are just one example currently under development. For these agents to be successful, they must be able to adapt to diverse conditions and learn continuously. Insights gained from the study of continuous learning in biological agents, including the use of multiple learning systems that operate in parallel, and that are optimized to learning in different environments, may be useful for developing more effective artificial agents. In addition, biological systems have decomposed the RL problem into sensory processing, value update and action output components. This allows the brain to optimize processing to the timescales of plasticity necessary for each system.

Most of the work in biological systems is based on either simple Pavlovian conditioning paradigms, which do not require an overt behavioural response, or two-armed bandit tasks in which animals have to learn which action is most valuable and update those action values as they change over time. Although Pavlovian conditioning and bandit learning are fundamental to much of the learning by biological systems, real behaviour in natural environments is much more complex. These problems are being addressed in artificial systems using hierarchical reinforcement learning. Many of the algorithms developed for studying these problems in artificial systems may be useful in biological systems. In addition, one of the difficult problems with HRL is learning how to decompose complex problems into sub-goals. At a behavioural level, biological systems do this routinely, and therefore insights from the study of behaviour in biological systems may translate to algorithms in artificial systems. Correspondingly, algorithms developed for artificial systems can help frame problems in biological experiments.

Understanding how the multiple neural systems in the brain can give rise to ongoing learning in diverse environments is already inspiring solutions for complex engineering problems, in the form

of novel algorithms and brain-inspired, neuromorphic hardware that can implement large spiking neural networks. Neuromorphic hardware operates on similar dynamical and architectural constraints as the brain, and thus provides an appealing platform for evaluating neuroscience-inspired solutions. Recently developed neuromorphic hardware tools are emerging as ideal candidates for real-world tasks on mobile platforms, thanks to their continuous inference and learning, which occur on an extremely tight energy budget.

Most state-of-the-art algorithms for RL take a domain general, or generalized function approximation approach and require vast amounts of data and training time. Decomposing the problem into state inference, value updating and action selection components, as is done in the brain, may allow for more efficient learning and the ability to track changes in the environment on fast timescales, similar to biological systems. Ongoing work at the interface of biological and artificial agents capable of reinforcement learning will provide deeper insights into the brain, and more effective artificial agents for solving real-world problems.

Received: 6 September 2018; Accepted: 16 January 2019;
Published online: 04 March 2019

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, 1998).
2. Pribram, K. H. A review of theory in physiological psychology. *Annu. Rev. Psychol.* **11**, 1–40 (1960).
3. Janak, P. H. & Tye, K. M. From circuits to behaviour in the amygdala. *Nature* **517**, 284–292 (2015).
4. Namburi, P. et al. A circuit mechanism for differentiating positive and negative associations. *Nature* **520**, 675–678 (2015).
5. Paton, J. J., Belova, M. A., Morrison, S. E. & Salzman, C. D. The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature* **439**, 865–870 (2006).
6. Hamid, A. A. et al. Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
7. Costa, V. D., Dal Monte, O., Lucas, D. R., Murray, E. A. & Averbeck, B. B. Amygdala and ventral striatum make distinct contributions to reinforcement learning. *Neuron* **92**, 505–517 (2016).
8. Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley, New York, 1994).
9. Bertsekas, D. P. *Dynamic Programming and Optimal Control* (Athena Scientific, Belmont, 1995).
10. Vapnik, V. *The Nature of Statistical Learning Theory* (Springer, New York, 2013).
11. Hessel, M. et al. Multi-task deep reinforcement learning with PopArt. Preprint at https://arxiv.org/abs/1809.04474 (2018).
12. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
13. Banino, A. et al. Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018).
14. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).
15. Rosenblatt, F. The perceptron: a probabilistic model for information-storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).
16. Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The "wake-sleep" algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).
17. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Current Research and Theory* (eds Black, A. H. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, New York, 1972).
18. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
19. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
20. Houk, J. C., Adamas, J. L. & Barto, A. G. in *Models of Information Processing in the Basal Ganglia* (eds Houk, J. C., Davis, J. L. & Beiser, D. G.) 249–274 (MIT Press, Cambridge, 1995).
21. Frank, M. J. Dynamic dopamine modulation in the basal ganglia: a neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *J. Cogn. Neurosci.* **17**, 51–72 (2005).
22. Haber, S. N., Kim, K. S., Mailly, P. & Calzavara, R. Reward-related cortical inputs define a large striatal region in primates that interface with associative cortical connections, providing a substrate for incentive-based learning. *J. Neurosci.* **26**, 8368–8376 (2006).
23. Mink, J. W. The basal ganglia: focused selection and inhibition of competing motor programs. *Prog. Neurobiol.* **50**, 381–425 (1996).
24. Lau, B. & Glimcher, P. W. Value representations in the primate striatum during matching behavior. *Neuron* **58**, 451–463 (2008).
25. O'Doherty, J. et al. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
26. Averbeck, B. B. & Costa, V. D. Motivational neural circuits underlying reinforcement learning. *Nat. Neurosci.* **20**, 505–512 (2017).
27. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
28. Schultz, W. Dopamine reward prediction error coding. *Dialog. Clin. Neurosci.* **18**, 23–32 (2016).
29. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
30. Saunders, B. T., Richard, J. M., Margolis, E. B. & Janak, P. H. Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties. *Nat. Neurosci.* **21**, 1072–1083 (2018).
31. Sharpe, M. J. et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).
32. Averbeck, B. B., Sohn, J. W. & Lee, D. Activity in prefrontal cortex during dynamic selection of action sequences. *Nat. Neurosci.* **9**, 276–282 (2006).
33. Seo, M., Lee, E. & Averbeck, B. B. Action selection and action value in frontal-striatal circuits. *Neuron* **74**, 947–960 (2012).
34. Lee, E., Seo, M., Dal Monte, O. & Averbeck, B. B. Injection of a dopamine type 2 receptor antagonist into the dorsal striatum disrupts choices driven by previous outcomes, but not perceptual inference. *J. Neurosci.* **35**, 6298–6306 (2015).
35. Averbeck, B. B., Lehman, J., Jacobson, M. & Haber, S. N. Estimates of projection overlap and zones of convergence within frontal-striatal circuits. *J. Neurosci.* **34**, 9497–9505 (2014).
36. Rothenhoefer, K. M. et al. Effects of ventral striatum lesions on stimulus versus action based reinforcement learning. *J. Neurosci.* **37**, 6902–6914 (2017).
37. Friedman, D. P., Aggleton, J. P. & Saunders, R. C. Comparison of hippocampal, amygdala, and perirhinal projections to the nucleus accumbens: combined anterograde and retrograde tracing study in the Macaque brain. *J. Comp. Neurol.* **450**, 345–365 (2002).
38. Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* **9**, 357–381 (1986).
39. Averbeck, B. B. Amygdala and ventral striatum population codes implement multiple learning rates for reinforcement learning. In *IEEE Symposium Series on Computational Intelligence* (IEEE, 2017).
40. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive mixtures of local experts. *Neural Comput.* **3**, 79–87 (1991).
41. Pfister, J. P., Toyoizumi, T., Barber, D. & Gerstner, W. Optimal spike-timing-dependent plasticity for precise action potential firing in supervised learning. *Neural Comput.* **18**, 1318–1348 (2006).
42. Benna, M. K. & Fusi, S. Computational principles of biological memory. Preprint at https://arxiv.org/abs/1507.07580 (2015).
43. Lahiri, S. & Ganguli, S. A memory frontier for complex synapses. In *Advances in Neural Information Processing Systems* Vol. 26 (eds Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z. & Weinberger, K. Q.) 1034–1042 (NIPS, 2013).
44. Koutnik, J., Greff, K., Gomez, F. & Schmidhuber, J. A clockwork RNN. Preprint at https://arxiv.org/abs/1402.3511 (2014).
45. Neil, D., M., P. & Liu, S.-C. Phased LSTM: accelerating recurrent network training for long or event-based sequences. In *Advances in Neural Information Processing Systems* Vol. 29 (eds Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. & Garnett, R.) 3882–3890 (NIPS, 2016).
46. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
47. O'Reilly, R. C. & Frank, M. J. Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Comput.* **18**, 283–328 (2006).
48. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer, New York, 2006).
49. Bottou, L. & LeCun, Y. Large scale online learning. In *Advances in Neural Information Processing Systems* Vol. 16 (eds Thrun, S., Saul, L. K. & Schölkopf, B.) (NIPS, 2004).
50. McCloskey, M. & Cohen, N. J. in *Psychology of Learning and Motivation: Advances in Research and Theory* Vol. 24 (ed. Bower, G. H.) 109–165 (1989).
51. McClelland, J. L., McNaughton, B. L. & O'Reilly, R. C. 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**, 419–457 (1995).
52. Kumaran, D., Hassabis, D. & McClelland, J. L. What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* **20**, 512–534 (2016).
53. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
54. Lin, L.-J. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Mach. Learn.* **8**, 293–321 (1992).
55. Zenke, F., Poole, B. & Ganguli, S. Continual learning through synaptic intelligence. Preprint at https://arxiv.org/abs/1703.04200 (2017).
56. Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M. & Tuytelaars, T. Memory aware synapses: learning what (not) to forget. Preprint at https://arxiv.org/abs/1711.09601 (2017).
57. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
58. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Reversal learning and dopamine: a Bayesian perspective. *J. Neurosci.* **35**, 2407–2416 (2015).
59. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
60. Glascher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
61. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
62. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine enhances model-based over model-free choice behavior. *Neuron* **75**, 418–424 (2012).
63. Miller, E. K. The prefrontal cortex and cognitive control. *Nat. Rev. Neurosci.* **1**, 59–65 (2000).
64. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).
65. Deserno, L. et al. Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc. Natl Acad. Sci. USA* **112**, 1595–1600 (2015).
66. Harlow, H. F. The formation of learning sets. *Psychol. Rev.* **56**, 51–65 (1949).
67. Iversen, S. D. & Mishkin, M. Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. *Exp. Brain Res.* **11**, 376–386 (1970).
68. Jang, A. I. et al. The role of frontal cortical and medial-temporal lobe brain areas in learning a Bayesian prior belief on reversals. *J. Neurosci.* **35**, 11751–11760 (2015).
69. Wang, J. X. et al. Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
70. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
71. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron* **91**, 1402–1412 (2016).
72. DeGroot, M. H. *Optimal Statistical Decisions* (Wiley, Hoboken, 1970).
73. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589 (2017).
74. Starkweather, C. K., Gershman, S. J. & Uchida, N. The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron* **98**, 616–629 (2018).
75. Wang, X. J. Synaptic basis of cortical persistent activity: the importance of NMDA receptors to working memory. *J. Neurosci.* **19**, 9587–9603 (1999).
76. Schöner, G. in *The Cambridge Handbook of Computational Psychology* (ed. Sun, R.) 101–126 (Cambridge Univ. Press, Cambridge, 2008).
77. Averbeck, B. B. Theory of choice in bandit, information sampling and foraging tasks. *PLoS Comput. Biol.* **11**, e1004164 (2015).
78. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proc. Natl Acad. Sci. USA* **110**, 20941–20946 (2013).
79. Akam, T., Costa, R. & Dayan, P. Simple plans or sophisticated habits? State, transition and learning interactions in the two-step task. *PLoS. Comput. Biol.* **11**, e1004648 (2015).
80. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
81. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**, 229–256 (1992).
82. Tesauro, G. Temporal difference learning and TD-Gammon. *Commun. ACM* **38**, 58–68 (1995).
83. Pomerleau, D. A. ALVINN: An autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems* Vol. 1 (ed. Touretzky, D. S.) (NIPS, 1988).
84. Levine, S., Finn, C., Darrell, T. & Abbeel, P. End-to-end training of deep visuomotor policies. *J. Mach. Learn. Res.* **17**, 1334–1373 (2016).
85. Gu, S., Lillicrap, T., Sutskever, I. & Levine, S. Continuous deep Q-learning with model-based acceleration. In *Proc. 33rd International Conference on Machine Learning* Vol. 48 2829–2838 (PMLR, 2016).
86. Burda, Y., Edwards, H., Storkey, A. & Klimov, O. Exploration by random network distillation. Preprint at https://arxiv.org/abs/1810.12894 (2018).
87. Vezhnevets, A. S. et al. Feudal networks for hierarchical reinforcement learning. Preprint at https://arxiv.org/abs/1703.01161 (2017).
88. Neftci, E. O. Data and power efficient intelligence with neuromorphic learning machines. *iScience* **5**, 52–68 (2018).
89. Kaiser, J., Mostafa, H. & Neftci, E. O. Synaptic plasticity dynamics for deep continuous local learning. Preprint at https://arxiv.org/abs/1811.10766 (2018).
90. Neftci, E. O., Augustine, C., Paul, S. & Detorakis, G. Event-driven random back-propagation: enabling neuromorphic deep learning machines. *Front. Neurosci.* **11**, 324 (2017).
91. Lillicrap, T. P., Cownden, D., Tweed, D. B. & Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* **7**, 13276 (2016).
92. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J. & Quillen, D. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *Int. J. Robot. Res.* **37**, 421–436 (2018).
93. Blundell, C. et al. Model-free episodic control. Preprint at https://arxiv.org/abs/1606.04460 (2016).
94. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Ann. Rev. Psychol.* **68**, 101–128 (2017).
95. Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. Preprint at https://arxiv.org/abs/1703.03400 (2017).
96. Ha, D. & Schmidhuber, J. World models. Preprint at https://arxiv.org/abs/1803.10122 (2018).
97. Zambaldi, V. et al. Relational deep reinforcement learning. Preprint at https://arxiv.org/abs/1806.01830 (2018).
98. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
99. Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vision Sci.* **1**, 417–446 (2015).
100. Bernacchia, A., Seo, H., Lee, D. & Wang, X. J. A reservoir of time constants for memory traces in cortical neurons. *Nat. Neurosci.* **14**, 366–372 (2011).
101. Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H. & Rushworth, M. F. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron* **65**, 927–939 (2010).
102. Iglesias, S. et al. Hierarchical prediction errors in midbrain and basal forebrain during sensory learning. *Neuron* **80**, 519–530 (2013).
103. Badre, D. & Frank, M. J. Mechanisms of hierarchical reinforcement learning in cortico-striatal circuits 2: evidence from fMRI. *Cereb. Cortex* **22**, 527–536 (2012).
104. Frank, M. J. & Badre, D. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb. Cortex* **22**, 509–526 (2012).
105. Botvinick, M. M. Hierarchical models of behavior and prefrontal function. *Trends Cogn. Sci.* **12**, 201–208 (2008).
106. Botvinick, M. M., Niv, Y. & Barto, A. C. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
107. Ribas-Fernandes, J. J. et al. A neural signature of hierarchical reinforcement learning. *Neuron* **71**, 370–379 (2011).
108. Botvinick, M. M. Hierarchical reinforcement learning and decision making. *Curr. Opin. Neurobiol.* **22**, 956–962 (2012).
109. Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. *Philos. Trans. R. Soc. Lond. B* **369**, 20130480 (2014).
110. Dayan, P. & Hinton, G. E. Feudal reinforcement learning. In *Advances in Neural Information Processing Systems* Vol. 5 (eds Hanson, S. J., Cowan, J. D. & Giles, C. L.) 271–278 (NIPS, 1992).
111. Sutton, R. S., Precup, D. & Singh, S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artif. Intell.* **112**, 181–211 (1999).
112. Bacon, P. L., Harb, J. & Precup, D. The option-critic architecture. *Proc. Thirty-First AAAI Conference on Artificial Intelligence* 1726–1734 (AAAI, 2017).
113. Jaderberg, M. et al. Reinforcement learning with unsupervised auxiliary tasks. Preprint at https://arxiv.org/abs/1611.05397 (2016).
114. Friston, K. The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).

115. Ross, S., Gordon, G. J. & Bagnell, J. A. A reduction of imitation learning and structured prediction to no-regret online learning. Preprint at https://arxiv.org/abs/1011.0686 (2010).
116. Le, H. M. et al. Hierarchical imitation and reinforcement learning. Preprint at https://arxiv.org/abs/1803.00590 (2018).
117. Koechlin, E., Ody, C. & Kouneiher, F. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).
118. Badre, D. & D'Esposito, M. Functional magnetic resonance imaging evidence for a hierarchical organization of the prefrontal cortex. *J. Cogn. Neurosci.* **19**, 2082–2099 (2007).
119. Muessgens, D., Thirugnanasambandam, N., Shitara, H., Popa, T. & Hallett, M. Dissociable roles of preSMA in motor sequence chunking and hand switching—a TMS study. *J. Neurophysiol.* **116**, 2637–2646 (2016).
120. Sabour, S., Frosst, N. & Hinton, G. E. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems* Vol. 30 (eds Guyon, I. et al.) 3856–3866 (2017).
121. Davies, M. et al. Loihi: a neuromorphic manycore processor with on-chip learning. *IEEE Micro* **38**, 82–99 (2018).
122. Friedmann, S. & Schemmel, J. Demonstrating hybrid learning in a flexible neuromorphic hardware system. Preprint at https://arxiv.org/abs/1604.05080 (2016).
123. Qiao, N. et al. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. *Front. Neurosci.* **9**, 141 (2015).
124. Neftci, E. et al. Synthesizing cognition in neuromorphic electronic systems. *Proc. Natl Acad. Sci. USA* **110**, 3468–3476 (2013).
125. Friedmann, S., Fremaux, N., Schemmel, J., Gerstner, W. & Meier, K. Reward-based learning under hardware constraints-using a RISC processor embedded in a neuromorphic substrate. *Front. Neurosci.* **7**, 160 (2013).
126. Hassabis, D., Kumaran, D., Summerfield, C. & Botvinick, M. Neuroscience-inspired artificial intelligence. *Neuron* **95**, 245–258 (2017).
127. Courbariaux, M., Bengio, Y. & David, J.-P. Training deep neural networks with low precision multiplications. Preprint at https://arxiv.org/abs/1412.7024 (2014).
128. Detorakis, G. et al. Neural and synaptic array transceiver: a brain-inspired computing framework for embedded learning. *Front. Neurosci.* **12**, 583 (2018).
129. Liu, S. C. & Delbruck, T. Neuromorphic sensory systems. *Curr. Opin. Neurobiol.* **20**, 288–295 (2010).

Acknowledgements

This work was supported by the Intramural Research Program of the National Institute of Mental Health (ZIA MH002928-01), and by the National Science Foundation under grant 1640081.

Competing interests

The authors declare no competing interests.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to B.B.A.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.