Neuron

CellPress

Review

Deep Reinforcement Learning and Its Neuroscientific Implications

Matthew Botvinick,[1,2,*] Jane X. Wang,[1] Will Dabney,[1] Kevin J. Miller,[1,2] and Zeb Kurth-Nelson[1,2]
[1]DeepMind, London, UK
[2]University College London, London, UK
*Correspondence: botvinick@google.com
https://doi.org/10.1016/j.neuron.2020.06.014

The emergence of powerful artificial intelligence (AI) is defining new research directions in neuroscience. To date, this research has focused largely on deep neural networks trained using supervised learning in tasks such as image classification. However, there is another area of recent AI work that has so far received less attention from neuroscientists but that may have profound neuroscientific implications: deep reinforcement learning (RL). Deep RL offers a comprehensive framework for studying the interplay among learning, representation, and decision making, offering to the brain sciences a new set of research tools and a wide range of novel hypotheses. In the present review, we provide a high-level introduction to deep RL, discuss some of its initial applications to neuroscience, and survey its wider implications for research on brain and behavior, concluding with a list of opportunities for next-stage research.

The past few years have seen a burst of interest in deep learning as a basis for modeling brain function (Cichy and Kaiser, 2019; Güçlü and van Gerven, 2017; Hasson et al., 2020; Marblestone et al., 2016; Richards et al., 2019). Deep learning has been studied for modeling numerous systems, including vision (Yamins et al., 2014; Yamins and DiCarlo, 2016), audition (Kell et al., 2018), motor control (Merel et al., 2019; Weinstein and Botvinick, 2017), navigation (Banino et al., 2018; Whittington et al., 2019), and cognitive control (Mante et al., 2013; Botvinick and Cohen, 2014). This resurgence of interest in deep learning has been catalyzed by recent dramatic advances in machine learning and artificial intelligence (AI). Of particular relevance is progress in training deep learning systems using supervised learning—that is, explicitly providing the "correct answers" during task training—on tasks such as image classification (Krizhevsky et al., 2012; Deng et al., 2009).

For all their freshness, the recent neuroscience applications of supervised deep learning can actually be seen as returning to a thread of research stretching back to the 1980s, when the first neuroscience applications of supervised deep learning began (Zipser and Andersen, 1988; Zipser, 1991). Of course this return is highly justified, given new opportunities that are presented by the availability of more powerful computers, allowing scaling of supervised deep learning systems to much more interesting datasets and tasks. However, at the same time, there are other developments in recent AI research that are more fundamentally novel and that have received less notice from neuroscientists. Our purpose in this review is to call attention to one such area that has vital implications for neuroscience, namely, deep reinforcement learning (RL).

As we will detail, deep RL brings deep learning together with a second computational framework that has already had a substantial impact on neuroscience research: RL. Although integrating RL with deep learning has been a long-standing aspiration in AI, it is only in very recent years that this integration has borne fruit. This engineering breakthrough has, in turn, brought to the fore a wide range of computational issues that do not arise within either deep learning or RL alone. Many of these relate in interesting ways to key aspects of brain function, presenting a range of inviting opportunities for neuroscientific research: opportunities that have so far been little explored.

In what follows, we start with a brief conceptual and historical introduction to deep RL and discuss why it is potentially important for neuroscience. We then highlight a few studies that have begun to explore the relationship between deep RL and brain function. Finally, we lay out a set of broad topics for which deep RL may provide new leverage for neuroscience, closing with a set of caveats and open challenges.

An Introduction to Deep RL

Reinforcement Learning

RL (Sutton and Barto, 2018) considers the problem of a learner or an agent embedded in an environment, where the agent must progressively improve the actions it selects in response to each environmental situation or state (Figure 1A). Critically, in contrast to supervised learning, the agent does not receive explicit feedback directly indicating correct actions. Instead, each action elicits a signal of associated reward or lack of reward, and the RL problem is to progressively update behavior so as to maximize the reward accumulated over time. Because the agent is not told directly what to do, it must explore alternative actions, accumulating information about the outcomes they produce, thereby gradually homing in on a reward-maximizing behavioral policy.

Note that RL is defined in terms of the learning problem, rather than by the architecture of the learning system or the learning algorithm itself. Indeed, a wide variety of architectures and algorithms have been developed, spanning a range of assumptions concerning what quantities are represented, how these are updated on the basis of experience, and how decisions are made.

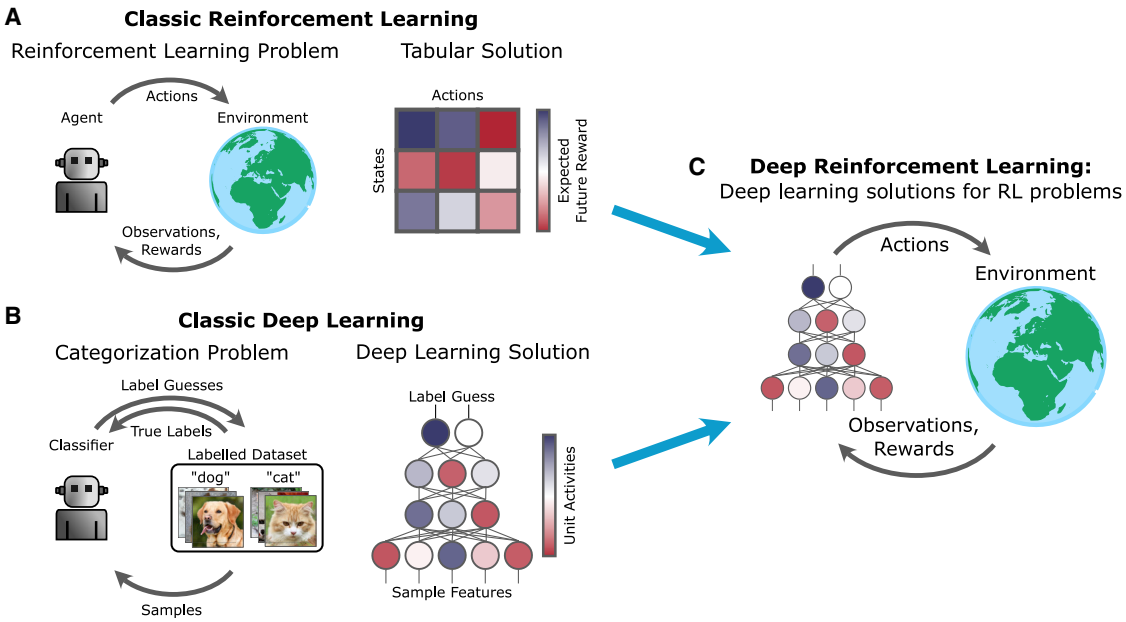

Figure 1. RL, Deep Learning, and Deep RL

(A) Left: the reinforcement learning problem. The agent selects actions and transmits them to the environment, which in turn transmits back to the agent observations and rewards. The agent attempts to select the actions that will maximize long-term reward. The best action might not result in immediate reward but might instead change the state of the environment to one in which reward can be obtained later. Right: tabular solution to a reinforcement learning problem. The agent considers the environment to be in one of several discrete states and learns from experience the expected long-term reward associated with taking each action in each state. These reward expectations are learned independently and do not generalize to new states or new actions.

(B) Left: the supervised learning problem. The agent receives a series of unlabeled data samples (e.g., images) and must guess the correct labels. Feedback on the correct label is provided immediately. Right: deep learning solution to a supervised learning problem. The features of a sample (e.g., pixel intensities) are passed through several layers of artificial neurons (circles). The activity of each neuron is a weighted sum of its inputs, and its output is a non-linear function of its activity. The output of the network is translated into a guess at the correct label for that sample. During learning, network weights are tuned such that these guesses come to approximate the true labels. These solutions have been found to generalize well to samples on which they have not been trained.

(C) Deep reinforcement learning, in which a neural network is used as an agent to solve a reinforcement learning problem. By learning appropriate internal representations, these solutions have been found to generalize well to new states and actions.

Fundamental to any solution of an RL problem is the question of how the state of the environment should be represented. Early work on RL involved simple environments comprising only a handful of possible states and simple agents that learned independently about each one, a so-called tabular state representation. By design, this kind of representation fails to support generalization—the ability to apply what is learned about one state to other similar states—a shortcoming that becomes increasingly inefficient as environments become larger and more complex, and individual states are therefore less likely to recur.

One important approach to attaining generalization across states is referred to as function approximation (Sutton and Barto, 2018), which attempts to assign similar representations to states in which similar actions are required. In one simple implementation of this approach, called linear function approximation, each state or situation is encoded as a set of features, and the learner uses a linear readout of these as a basis for selecting its actions.

Although linear function approximation has been often used in RL research, it has long been recognized that what is needed for RL to produce intelligent, human-like behavior is some form of non-linear function approximation. Just as recognizing visual categories (e.g., "cat") is well known to require non-linear processing of visual features (edges, textures, and more complex configurations), non-linear processing of perceptual inputs is generally required in order to decide on adaptive actions.

In acknowledgment of this point, RL research has long sought workable methods for non-linear function approximation. Although a variety of approaches have been explored over the years, often treating the representation learning problem independent of the underlying RL problem (Mahadevan and Maggioni, 2007; Konidaris et al., 2011), a long-standing aspiration has been to perform adaptive non-linear function approximation using deep neural networks.

Deep Learning

Deep neural networks are computational systems composed of neuron-like units connected through synapse-like contacts (Figure 1B). Each unit transmits a scalar value, analogous to a spike rate, which is computed on the basis of the sum of its inputs, that is, the activities of "upstream" units multiplied by the strength of the transmitting synapse or connection (Goodfellow et al., 2016). Critically, unit activity is a non-linear function of these inputs, allowing networks with layers of units interposed between the "input" and "output" sides of the system (i.e., "deep" neural networks) to approximate any function mapping activation inputs to activation outputs (Sutskever and Hinton, 2008). Furthermore, when the connectivity pattern includes loops, as in "recurrent" neural networks, the network's

activations can preserve information about past events, allowing the network to compute functions on the basis of sequences of inputs.

"Deep learning" refers to the problem of adjusting the connection weights in a deep neural network so as to establish a desired input-output mapping. Although a number of algorithms exist for solving this problem, by far the most efficient and widely used is backpropagation, which uses the chain rule from calculus to decide how to adjust weights throughout a network.

Although backpropagation was developed well over 30 years ago (Rumelhart et al., 1985; Werbos, 1974), until recently it was used almost exclusively for supervised learning, as defined above, or for unsupervised learning, in which only inputs are presented, and the task is to learn a "good" representation of those inputs on the basis of some function evaluating representational structure, as is done for example in clustering algorithms. Importantly, both of these learning problems differ fundamentally from RL. In particular, unlike supervised and unsupervised learning, RL requires exploration, as the learner is responsible for discovering actions that increase reward. Furthermore, exploration must be balanced against leveraging action-value information already acquired, or as it is conventionally put, exploration must be weighed against "exploitation." Unlike with most traditional supervised and unsupervised learning problems, a standard assumption in RL is that the actions of the learning system affect its inputs on the next time step, creating a sensory-motor feedback loop and potential difficulties due to nonstationarity in the training data. This creates a situation in which target behaviors or outputs involve multi-step decision processes rather than single input-output mappings. Until very recently, applying deep learning to RL settings has stood as a frustratingly impenetrable problem.

Deep Reinforcement Learning

Deep RL leverages the representational power of deep learning to tackle the RL problem. We define a deep RL system as any system that solves an RL problem (i.e., maximizes long-term reward), using representations that are themselves learned by a deep neural network (rather than stipulated by the designer). Typically, deep RL systems use a deep neural network to compute a non-linear mapping from perceptual inputs to action values (e.g., Mnih et al., 2015) or action probabilities (e.g., Silver et al., 2016), as well as RL signals that update the weights in this network, often via backpropagation, in order to produce better estimates of reward or to increase the frequency of highly rewarded actions (Figure 1C).

A notable early precursor to modern-day successes with deep RL occurred in the early 1990s, with a system nicknamed TD-Gammon, which combined neural networks with RL to learn how to play backgammon competitively with top human players (Tesauro, 1994). More specifically, TD-Gammon used a temporal difference RL algorithm, which computed an estimate for each encountered board position of how likely the system was to win (a state-value estimate). The system then computed a reward-prediction error (RPE)—essentially an indication of positive surprise or disappointment—on the basis of subsequent events. The RPE was fed as an error signal into the backpropagation algorithm, which updated the network's weights so as to yield more accurate state-value estimates. Actions could then be selected so as to maximize the state value for the next board state. In order to generate many games on which to train, TD-Gammon used self-play, in which the algorithm would play moves against itself until one side won.

Although TD-Gammon provided a tantalizing example of what RL implemented via neural networks might deliver, its approach yielded disappointing results in other problem domains. The main issue was instability; whereas in tabular and linear systems, RL reliably moved toward better and better behaviors, when combined with neural networks, the models often collapsed or plateaued, yielding poor results.

This state of affairs changed dramatically in 2013, with the report of the Deep Q Network (DQN), the first deep RL system that learned to play classic Atari video games (Mnih et al., 2013, 2015). Although DQN was widely noted for attaining better-than-human performance on many games, the real breakthrough was simply in getting deep RL to work in a reliably stable way. It incorporated several mechanisms that reduced nonstationarity, treating the RL problem more like a series of supervised learning problems, upon which the tools of deep learning could be more reliably applied. One example is "experience replay" (Lin, 1991), in which past state-action-reward-next-state transitions were stored away and intermittently re-presented in random order in order to mimic the random sampling of training examples that occurs in supervised learning. This helped greatly reduce variance and stabilize the updates.

Since DQN, work on deep RL has progressed and expanded at a remarkable pace. Deep RL has been scaled up to highly complex game domains ranging from Dota (Berner et al., 2019) to StarCraft II (Vinyals et al., 2019) to capture the flag (Jaderberg et al., 2019). Novel architectures have been developed that support effective deep RL in tasks requiring detailed long-term memory (Graves et al., 2016; Wayne et al., 2018). Deep RL has been integrated with model-based planning, resulting in superhuman play in complex games including chess and go (Silver et al., 2016, 2017a, 2017b, 2018). Furthermore, methods have been developed to allow deep RL to tackle difficult problems in continuous motor control, including simulations of soccer and gymnastics (Merel et al., 2018; Heess et al., 2016), and robotics problems such as in-hand manipulation of a Rubik's cube (Akkaya et al., 2019). We review some of these developments in greater detail below, as part of a larger consideration of what implications deep RL may have for neuroscience, the topic to which we now turn.

Deep RL and Neuroscience

Deep RL is built from components—deep learning and RL—that have already independently had a profound impact within neuroscience. Deep neural networks have proved to be an outstanding model of neural representation (Yamins et al., 2014; Sussillo et al., 2015; Kriegeskorte, 2015; Mante et al., 2013; Pandarinath et al., 2018; Rajan et al., 2016; Zipser, 1991; Zipser and Andersen, 1988; Figure 2A). However, this research has for the most part used supervised training and has therefore provided little direct leverage on the big-picture problem of understanding motivated, goal-directed behavior within a sensory-motor loop. At the same time, RL has provided a powerful theory of the neural mechanisms of learning and decision making (Niv, 2009). This

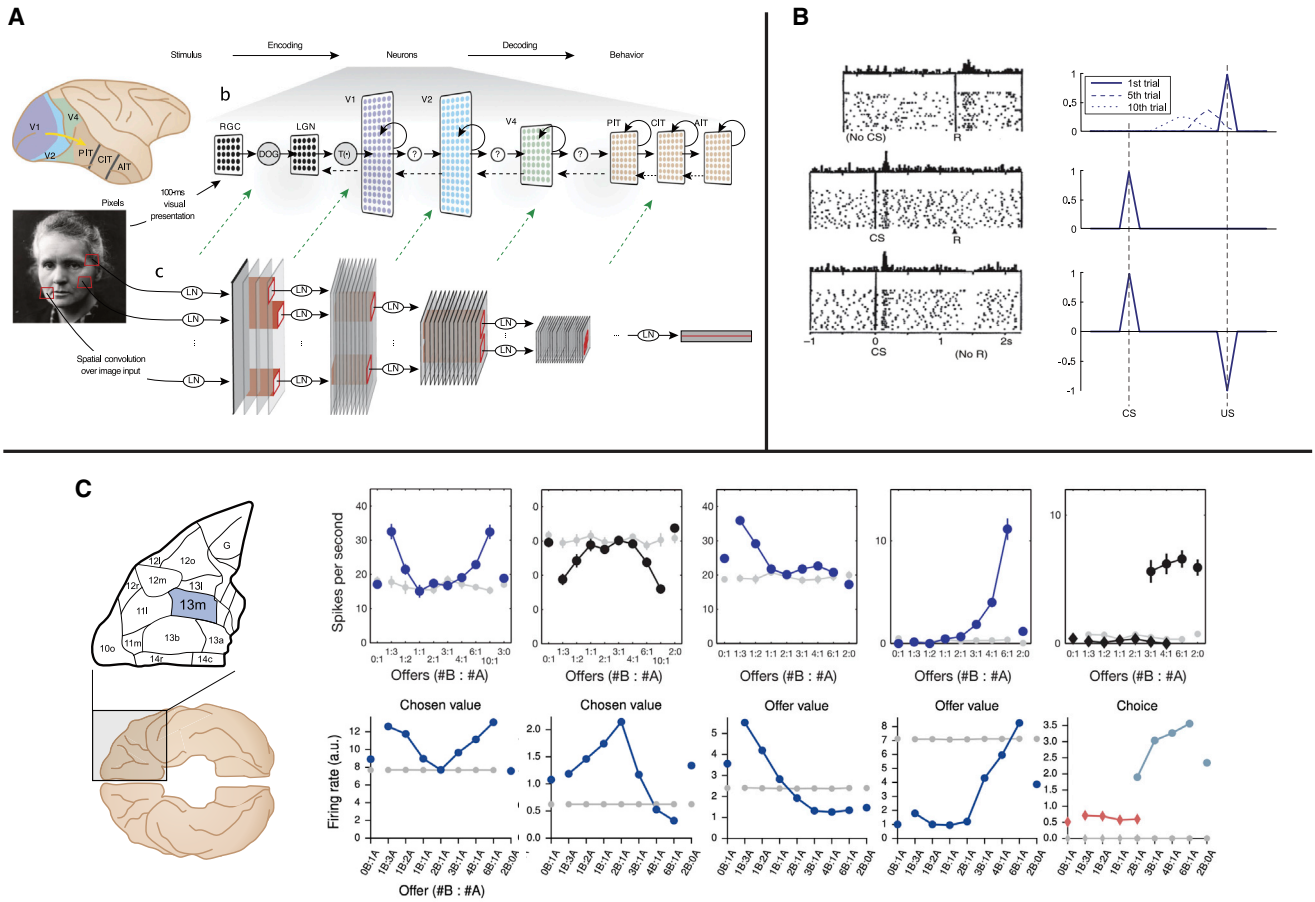

Figure 2. Applications to Neuroscience

(A) Supervised deep learning has been used in a wide range of studies to model and explain neural activity. In one influential study, Yamins and DiCarlo (2016) used a deep convolutional network (shown schematically in the lower portion of the figure) to model single-unit responses in various portions of the macaque ventral stream (upper portion). Figure adapted from Yamins and DiCarlo (2016).

(B) Reinforcement learning has been connected with neural function in a number of ways. Perhaps most influential has been the link established between phasic dopamine release and the temporal-difference reward-prediction error signal (RPE). The left side of the panel shows typical spike rasters and histograms from dopamine neurons in ventral tegmental area under conditions in which a food reward arrives unpredictably (top), arrives following a predictive cue (conditional stimulus [CS]), or is withheld following a CS. The corresponding panels on the right plot RPEs from a temporal-difference RL model under parallel conditions, showing qualitatively identical dynamics. Figure adapted from Niv (2009).

(C) Applications of deep RL to neuroscience have only just begun. In one pioneering study, Song et al. (2017) trained a recurrent deep RL network on a reward-based decision-making task paralleling one that had been studied in monkeys by Padoa-Schioppa and Assad (2006). The latter study examined the responses of neurons in orbitofrontal area 13 m (see left panel) across many different choice sets involving two flavors of juice in particular quantities (x axes in upper plots), reporting neurons whose activity tracked the inferred value of the monkey's preferred choice (two top left panels), the value of each individual juice (next two panels), or the identity of the juice actually chosen (right panel). Examining units within their deep RL model, Song et al. (2017) found patterns of activity closely resembling the neurophysiological data (bottom panels). Panels adapted from Song et al. (2017), Padoa-Schioppa and Assad (2006), and Stalnaker et al. (2015).

theory most famously explains the activity of dopamine neurons as a RPE (Watabe-Uchida et al., 2017; Glimcher, 2011; Lee et al., 2012; Daw and O'Doherty, 2014; Figure 2B) but also accounts for the role of a wide range of brain structures in reward-driven learning and decision making (Stachenfeld et al., 2017; Botvinick et al., 2009; O'Reilly and Frank, 2006; Gläscher et al., 2010; Wang et al., 2018; Wilson et al., 2014b). It has been integrated into small neural networks with handcrafted structure to provide models of how multiple brain regions may interact to guide learning and decision-making (O'Reilly and Frank, 2006; Frank and Claus, 2006). Just as in the machine learning context, however, RL itself has until recently offered neuroscience little guidance in thinking about the problem of representation (for discussion, see Botvinick et al., 2015; Wilson et al., 2014b; Stachenfeld et al., 2017; Behrens et al., 2018; Gershman et al., 2010).

Deep RL offers neuroscience something new, by showing how RL and deep learning can fit together. While deep learning focuses on how representations are learned, and RL on how rewards guide learning, in deep RL new phenomena emerge: processes by which representations support, and are shaped by, reward-driven learning and decision making.

If deep RL offered no more than a concatenation of deep learning and RL in their familiar forms, it would be of limited import. But deep RL is more than this; when deep learning and RL are integrated, each triggers new patterns of behavior in the other, leading to computational phenomena unseen in either

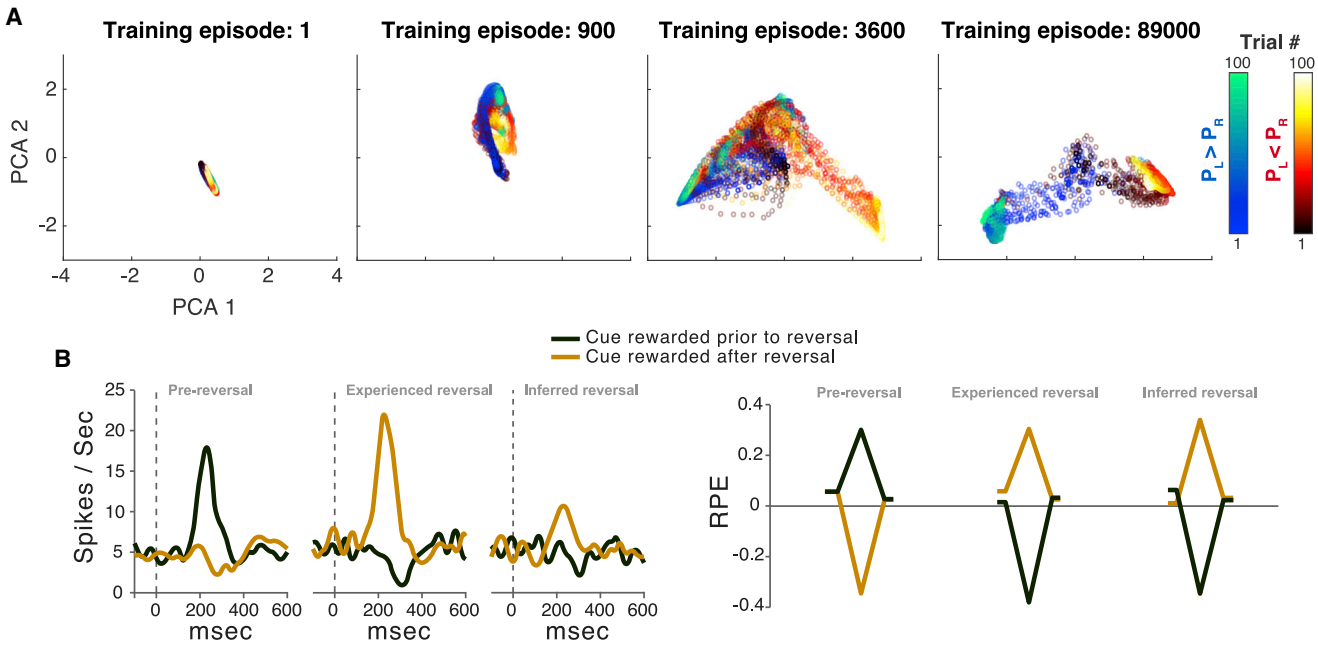

Figure 3. Meta-Reinforcement Learning

(A) Visualization of representations learned through meta-reinforcement learning, at various stages of training. An artificial agent is trained on a series of independent Bernoulli two-armed bandits (100 trials per episode), such that the probabilities of reward payout P_L and P_R are drawn uniformly from U(0, 1). Scatter points depict the first two principal components of the recurrent neural network (RNN) activation (LSTM output) vector taken from evaluation episodes at certain points in training, colored according to trial number (darker, earlier trials) and whether $P_L > P_R$. Only episodes for which $|P_L - P_R| > 0.3$ are plotted are shown. (B) Panels adapted from Bromberg-Martin et al. (2010) and Wang et al. (2018). Left: dopaminergic activity in response to cues ahead of a reversal and for cues with an experienced and inferred change in value. Right: corresponding RPE signals from an artificial agent. Leading and trailing points for each data series correspond to initial fixation and saccade steps. Peaks and troughs correspond to stimulus presentation.

deep learning or RL on their own. That is to say, deep RL is much more than the sum of its parts. And the novel aspects of the integrated framework in turn translate into new explanatory principles, hypotheses, and available models for neuroscience.

We unpack this point in the next section in considering some of the few neuroscience studies to have appeared so far that have leveraged deep RL, turning subsequently to a consideration of some wider issues that deep RL raises for neuroscience research.

Vanguard Studies

Although a number of commentaries have appeared that address aspects of deep RL from a neuroscientific perspective (Hassabis et al., 2017; Zador, 2019; Marblestone et al., 2016), few studies have yet applied deep RL models directly to neuroscientific data.

In a few cases, researchers have deployed deep RL in ways analogous to previous applications of supervised deep learning and RL. For example, transplanting a long-standing research strategy from deep learning (Yamins et al., 2014; Zipser, 1991) to deep RL, Song et al. (2017) trained a recurrent deep RL model on a series of reward-based decision making tasks that have been studied in the neuroscience literature, reporting close correspondences between the activation patterns observed in the network's internal units and neurons in dorsolateral prefrontal, orbitofrontal, and parietal cortices (Figure 2C). Work by Banino et al. (2018) combined supervised deep learning and deep RL methods to show how grid-like representations resembling those seen in entorhinal cortex can enhance goal-directed navigation performance.

As we have stressed, phenomena arise within deep RL that do not arise in deep learning or RL considered separately. A pair of recent studies focused on the neuroscientific implications of these emergent phenomena. In one, Wang et al. (2018) examined the behavior of recurrent deep RL systems and described a novel meta-RL effect: when trained on a series of interrelated tasks (e.g., a series of forced-choice decision tasks with the same overall structure but different reward probabilities) recurrent deep RL networks develop the ability to adapt to new tasks of the same kind without weight changes. This is accompanied by correspondingly structured representations in the activity dynamics of the hidden units that emerge throughout training (Figure 3A). Slow RL-driven learning at the level of the network's connection weights shape the network's activation dynamics such that rapid behavioral adaptation can be driven by those activation dynamics alone, akin to the idea from neuroscience that RL can be supported, in some cases, by activity-based working memory (Collins and Frank, 2012). In short, slow RL spontaneously gives rise to a separate and faster RL algorithm. Wang et al. (2018) showed how this meta-RL effect could be applied to explain a wide range of previously puzzling findings from neuroscientific studies of dopamine and prefrontal cortex function (Figure 3B).

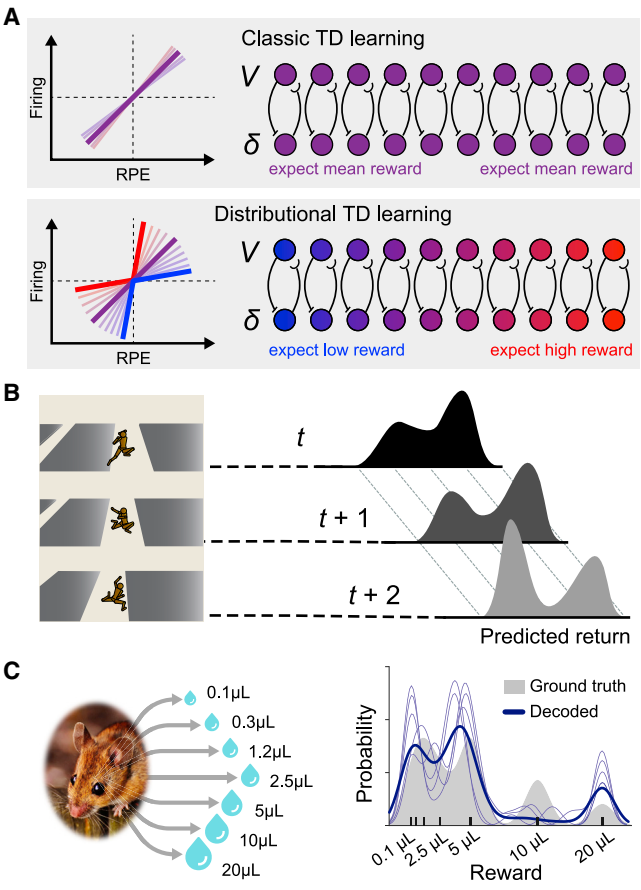

Figure 4. Distributional RL

(A) Top: in the classic temporal difference (TD) model, each dopamine cell computes a prediction error (δ) with respect to the same predicted reward (V). Bottom: in distributional TD, some RPE channels amplify negative RPEs (blue) and others amplify positive RPEs (red). This causes the channels to learn different reward predictions, ranging from very pessimistic (blue) to very optimistic (red).

(B) Artificial agents endowed with diverse RPE scaling learn to predict the return distribution. In this example, the agent is uncertain whether it will successfully land on the platform. The agent's predicted reward distribution on three consecutive time steps is shown at right.

(C) In real animals, it is possible to decode the reward distribution directly from dopamine activity. Here, mice were extensively trained on a task with probabilistic reward. The actual reward distribution of the task is shown as a gray shaded area. When interpreted as RPE channels of a distributional TD learner, the firing of dopamine cells decodes to the distribution shown in blue (thin traces are the best five solutions and the thick trace is their mean). The decoded distribution matches multiple modes of the actual reward distribution. Panels adapted from Dabney et al. (2020).

A second such study was conducted by Dabney et al. (2020). They leveraged a deep RL technique developed in recent AI work and referred to as distributional RL (Bellemare et al., 2017). Earlier, in discussing the history of deep RL, we mentioned the RPE. In conventional RL, this signal is a simple scalar, with positive numbers indicating a positive surprise and negative ones indicating disappointment. More recent neuroscientifically inspired models have suggested that accounting for the distribution and uncertainty of reward is important for decision making under risk (Mikhael and Bogacz, 2016). In distributional RL, the RPE is expanded to a vector, with different elements signaling RPE signals on the basis of different a priori forecasts, ranging from highly optimistic to highly pessimistic predictions (Figures 4A and 4B). This modification had been observed in AI work to dramatically enhance both the pace and outcome of RL across a variety of tasks, something, importantly, that is observed in deep RL but not in simpler forms such as tabular or linear RL (in part because of the impact of distributional coding on representation learning; Lyle et al., 2019). Carrying this finding into the domain of neuroscience, Dabney et al. (2020) studied electrophysiological data from mice to test whether the dopamine system might use the kind of vector code involved in distributional RL. As noted earlier, dopamine has been proposed to transmit an RPE-like signal. Dabney et al. (2020) obtained strong evidence that this dopaminergic signal is distributional, conveying a spectrum of RPE signals ranging from pessimistic to optimistic (Figure 4C).

Topics for Next-Step Research

As we have noted, explorations of deep RL in neuroscience have only just begun. What are the key opportunities going forward? In the sections below we outline six areas where it appears that deep RL may provide leverage for neuroscientific research. In each case, intensive explorations are already under way in the AI context, providing neuroscience with concrete opportunities for translational research. Although we stress tangible proposals in what follows, it is important to bear in mind that these proposals do not restrict the definition of deep RL. Deep RL is instead a broad and multi-faceted framework, within which algorithmic details can be realized in a huge number of ways, making the space of resulting hypotheses for neuroscience bracingly diverse.

Representation Learning

The question of representation has long been central to neuroscience, beginning perhaps with the work of Hubel and Weisel (1959) and continuing robustly to the present day (Constantinescu et al., 2016; Stachenfeld et al., 2017; Wilson et al., 2014b). Neuroscientific studies of representation have benefited from tools made available by deep learning (Zipser and Andersen, 1988; Yamins et al., 2014), which provides models of how representations can be shaped by sensory experience. Deep RL expands this toolkit, providing for the first time models of how representations can be shaped by rewards and by task demands. In a deep RL agent, reward-based learning shapes internal representations, and these representations in turn support reward-based decision making. A canonical example would be the DQN network training on an Atari task. Here, reward signals generated on the basis of how many points are scored feed into a backpropagation algorithm that modifies weights throughout the deep neural network, updating the response profiles of all units. This results in representations that are appropriate for the task. Whereas a supervised learning system assigns similar representations to images with similar labels (Figures 5A and 5B), deep RL tends to associate images with similar functional task implications (Figures 5C and 5D).

This idea of reward-based representation learning resonates with a great deal of evidence from neuroscience. We know, for example, that representations of visual stimuli in prefrontal

cortex depend on which task an animal has been trained to perform (Freedman et al., 2001) and that effects of task reward on neural responses can be seen even in primary visual cortex (Pakan et al., 2018).

The development and use of deep RL systems has raised awareness of two serious drawbacks of representations that are shaped by RL alone. One problem is that task-linked rewards are generally sparse. In chess, for example, reward occurs once per game, making it a weak signal for learning about opening moves. A second problem is overfitting: internal representations shaped exclusively by task-specific rewards may end up being useful only for tasks the learner has performed but completely wrong for new tasks (Zhang et al., 2018; Cobbe et al., 2019). Better would be some learning procedure that gives rise to internal representations that are more broadly useful, supporting transfer between tasks.

To address these issues, deep RL is often supplemented in practice with either unsupervised learning (Higgins et al., 2017), or "self-supervised" learning. In self-supervised learning the agent is trained to produce, in addition to an action, some auxiliary output that matches a training signal that is naturally available from the agent's stream of experience, regardless of what specific RL task it is being trained on (Jaderberg et al., 2016; Banino et al., 2018). An example is prediction learning, in which the agent is trained to predict, on the basis of its current situation, what it will observe at future time steps (Wayne et al., 2018; Gelada et al., 2019). Unsupervised and self-supervised learning mitigate both problems associated with pure RL, as they shape representations in a way that is not tied exclusively to the specific tasks confronted by the learner, thus yielding representations that have the potential to support transfer to other tasks when they arise. All of this is consistent with existing work in neuroscience, in which unsupervised learning (e.g., Olshausen and Field, 1996; Hebb, 1949; Kohonen, 2012) and prediction learning (e.g., Schapiro et al., 2013; Stachenfeld et al., 2017; Rao and Ballard, 1999) have been proposed to shape internal representations. Deep RL offers the opportunity to pursue these ideas in a setting in which these forms of learning can mix with reward-driven learning (Marblestone et al., 2016; Richards et al., 2019) and the representations they produce support adaptive behavior.

One further issue foregrounded in deep RL involves the role of inductive biases in shaping representation learning. Most deep RL systems that take visual inputs use a processing architecture (a convolutional network; Fukushima, 1980) that biases them toward representations that take into account the translational invariance of images. And more recently developed architectures build in a bias to represent visual inputs as comprising sets of discrete objects with recurring pairwise relationships (Watters et al., 2019; Battaglia et al., 2018). Such ideas recall existing neuroscientific findings (Roelfsema et al., 1998) and have interesting consequences in deep RL, such as the possibility of exploring and learning much more efficiently by decomposing the environment into objects (Diuk et al., 2008; Watters et al., 2019).

Model-Based RL

An important classification of RL algorithms is between "model-free" algorithms, which learn a direct mapping from perceptual inputs to action outputs, and "model-based" algorithms, which instead learn a "model" of action-outcome relationships and use this to plan actions by forecasting their outcomes.

This dichotomy has had a marked impact in neuroscience, in which brain regions have been accorded different roles in these two kinds of learning, and an influential line of research has focused on how the two forms of learning may trade off against each other (Lee et al., 2014; Daw et al., 2005, 2011; Balleine and Dickinson, 1998; Dolan and Dayan, 2013). Deep RL opens up a new vantage point on the relationship between model-free and model-based RL. For example, in AlphaGo and its successor systems (Silver et al., 2016, 2017b, 2018), model-based planning is guided in part by value estimates and action tendencies learned through model-free RL. Related interactions between the two systems have been studied in neuroscience and psychology (Cushman and Morris, 2015; Keramati et al., 2016).

In AlphaGo, the action-outcome model used in planning is hand coded. Still more interesting from a neuroscientific point of view (Gläscher et al., 2010) is recent work in which model-based RL relies on models learned from experience (Schrittwieser et al., 2019; Nagabandi et al., 2018; Ha and Schmidhuber, 2018). Although these algorithms have achieved great success in some domains, a key open question is whether systems can learn to capture transition dynamics at a high level of abstraction ("If I throw a rock at that window, it will shatter") rather than being tied to detailed predictions about perceptual observations (predicting where each shard would fall) (Behrens et al., 2018; Konidaris, 2019).

One particularly intriguing finding from deep RL is that there are circumstances under which processes resembling model-based RL may emerge spontaneously within systems trained using model-free RL algorithms (Wang et al., 2016; Guez et al., 2019). The neuroscientific implications of this "model-free planning" have already been studied in a preliminary way (Wang et al., 2018), but it deserves further investigation. Intriguingly, model-based behavior is also seen in RL systems that use a particular form of predictive code, referred to as the "successor representation" (Vértes and Sahani, 2019; Momennejad, 2020; Kulkarni et al., 2016; Barreto et al., 2017), suggesting one possible mechanism through which model-free planning might arise.

An interesting question that has arisen in neuroscientific work is how the balance between model-free and model-based RL is arbitrated, that is, what are the mechanisms that decide, moment to moment, whether behavior is controlled by model-free or model-based mechanisms (Daw et al., 2005; Lee et al., 2014). Related to this question, some deep RL work in AI has introduced mechanisms that learn through RL whether and how deeply to plan before committing to an action (Hamrick et al., 2017). The resulting architecture is reminiscent of work from neuroscience on cognitive control mechanisms implemented in the prefrontal cortex (Botvinick and Cohen, 2014), a topic we discuss further below.

Memory

On the topic of memory, arguably one of the most important in neuroscience, deep RL once again opens up fascinating new questions and highlights novel computational possibilities. In particular, deep RL provides a computational setting in which

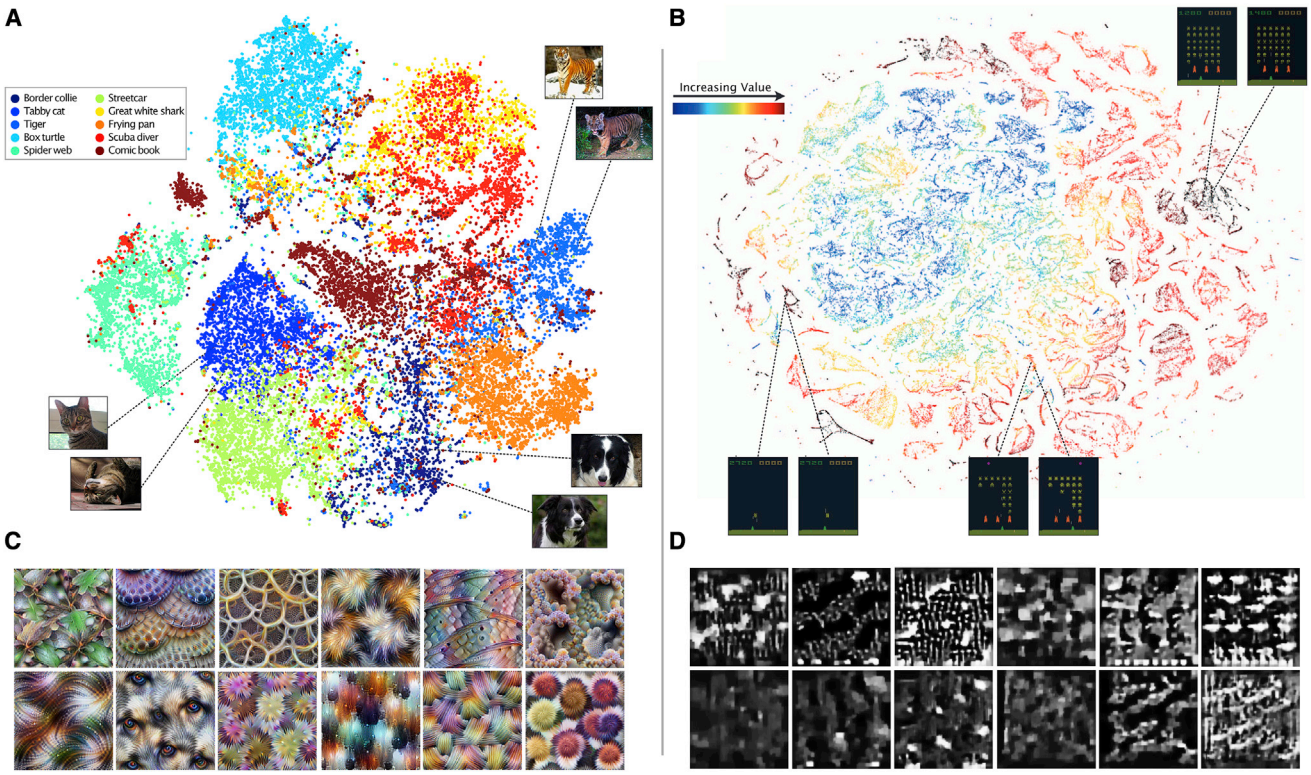

Figure 5. Representations Learned by Deep Supervised Learning and Deep RL

(A) Representations of natural images (ImageNet; Deng et al., 2009) from a deep neural network trained to classify objects (Carter et al., 2019). The t-distributed stochastic neighbor embedding (t-SNE) of representations in one layer ("mixed5b"), colored by predicted object class, and with example images shown.

(B) Synthesized inputs that maximally activate individual artificial neurons (in layer "mixed4a") show specialization for high-level features and textures to support object recognition (Olah et al., 2017).

(C) Representations of Atari video game images (Bellemare et al., 2013) from a DQN agent trained with deep RL (Mnih et al., 2015). The t-SNE of representations from the final hidden layer, colored by predicted future reward value, and with example images shown.

(D) Synthesized images that maximally activate individual cells from the final convolutional layer reveal texture-like detail for reward-predictive features (Such et al., 2019). For example, in the game Seaquest, the relative position of the submarine to incoming fish appears to be captured in the top rightmost image.

to investigate how memory can support reward-based learning and decision making, a topic that has been of growing interest in neuroscience (see, e.g., Eichenbaum et al., 1999; Gershman and Daw, 2017). The first broadly successful deep RL models relied on experience replay (Mnih et al., 2013), wherein past experiences are stored and intermittently used alongside new experiences to drive learning. This has an intriguing similarity to the replay events observed in hippocampus and elsewhere and indeed was inspired by this phenomenon and its suspected role in memory consolidation (Wilson and McNaughton, 1994; Kumaran et al., 2016). Although early deep RL systems replayed experience uniformly, replay in the brain is not uniform (Mattar and Daw, 2018; Gershman and Daw, 2017; Gupta et al., 2010; Carey et al., 2019), and non-uniformity has been explored in machine learning as a way to enhance learning (Schaul et al., 2015).

In addition to driving consolidation, memory maintenance and retrieval in the brain are also used for online decision making (Pfeiffer and Foster, 2013; Wimmer and Shohamy, 2012; Bornstein and Norman, 2017; O'Reilly and Frank, 2006). In deep RL, two kinds of memory serve this function. First, "episodic" memory systems read and write long-term storage slots (Wayne et al., 2018; Lengyel and Dayan, 2008; Blundell et al., 2016). One interesting aspect of these systems is that they allow relatively easy analysis of what information is being stored and retrieved at each time step (Graves et al., 2016; Banino et al., 2020), inviting comparisons with neural data. Second, recurrent neural networks store information in activations, in a manner similar to what is referred to in neuroscience as working memory maintenance. The widely used "LSTM" (long short-term memory) and "GRU" (gated recurrent unit) architectures use learnable gating to forget or retain task-relevant information, reminiscent of similar mechanisms that have been proposed to exist in the brain (Chatham and Badre, 2015; Stalter et al., 2020).

Still further deep RL memory mechanisms are being invented at a rapid rate, including systems that deploy attention and relational processing over information in memory (e.g., Parisotto et al., 2019; Graves et al., 2016) and systems that combine and coordinate working and episodic memory (e.g., Ritter et al., 2018). This represents one of the topic areas where an exchange between deep RL and neuroscience seems most actionable and most promising.

Exploration

As noted earlier, exploration is one of the features that differentiate RL from other standard learning problems. RL imposes the

need to seek information actively, testing out novel behaviors and balancing them against established knowledge, negotiating the explore-exploit trade-off. Animals, of course, face this challenge as well, and it has been of considerable interest in neuroscience and psychology (see, e.g., Costa et al., 2019; Gershman, 2018; Wilson et al., 2014a; Schwartenbeck et al., 2013). Here once again, deep RL offers a new computational perspective and a set of specific algorithmic ideas.

A key strategy in work on exploration in RL has been to include an auxiliary ("intrinsic") reward (Schmidhuber, 1991; Dayan and Balleine, 2002; Chentanez et al., 2005; Oudeyer et al., 2007), such as for novelty, which encourages the agent to visit unfamiliar states or situations. However, because deep RL generally deals with high-dimensional perceptual observations, it is rare for exactly the same perceptual observation to recur. The question thus arises of how to quantify novelty, and a range of innovative techniques have been proposed to address this problem (Bellemare et al., 2016; Pathak et al., 2017; Burda et al., 2019; Badia et al., 2020). Another approach to intrinsically motivated exploration is to base it not on novelty but on uncertainty, encouraging the agent to enter parts of the environment where its predictions are less confident (Osband et al., 2016). And still other work has pursued the idea of allowing agents to learn or evolve their own intrinsic motivations, on the basis of task experience (Niekum et al., 2010; Singh et al., 2010; Zheng et al., 2018).

Meta-RL provides another interesting and novel perspective on exploration. As noted earlier, meta-RL gives rise to activation dynamics that support learning, even when weight changes are suspended. Importantly, the learning that occurs in that setting involves exploration, which can be quite efficient because it is structured to fit with the kinds of problems the system was trained on. Indeed, exploration in meta-RL systems can look more like hypothesis-driven experimentation than random exploration (Denil et al., 2016; Dasgupta et al., 2019). These properties of meta-RL systems make them an attractive potential tool for investigating the neural basis of strategic exploration in animals.

Finally, some research in deep RL proposes to tackle exploration by sampling randomly in the space of hierarchical behaviors (Machado et al., 2017; Jinnai et al., 2020; Hansen et al., 2020). This induces a form of directed, temporally extended, random exploration reminiscent of some animal foraging models (Viswanathan et al., 1999).

Cognitive Control and Action Hierarchies

Cognitive neuroscience has long posited a set of functions, collectively referred to as "cognitive control," that guide task selection and strategically organize cognitive activity and behavior (Botvinick and Cohen, 2014). The very first applications of deep RL contained nothing corresponding to this set of functions. However, as deep RL research has developed, it has begun to grapple with the problem of attaining competence and switching among multiple tasks or skills, and in this context a number of computational techniques have been developed that bear an intriguing relationship with neuroscientific models of cognitive control.

Perhaps most relevant is research that has adapted to deep RL ideas originating from the older field of hierarchical RL. Here, RL operates at two levels, shaping a choice among high-level multi-step actions (e.g., "make coffee") and also among actions at a more atomic level (e.g., "grind beans"; see Botvinick et al., 2009). Deep RL research has adopted this hierarchical scheme in a number of ways (Bacon et al., 2017; Harutyunyan et al., 2019; Barreto et al., 2019; Vezhnevets et al., 2017). In some of these, the low-level system can operate autonomously, and the higher level system intervenes only at a cost that makes up part of the RL objective (Teh et al., 2017; Harb et al., 2018), an arrangement that resonates with the notions in neuroscience of habit pathways and automatic versus controlled processing (Dolan and Dayan, 2013; Balleine and O'Doherty, 2010), as well as the idea of a "cost of control" (Shenhav et al., 2017). In deep RL, the notion of top-down control over lower level habits has also been applied in motor control tasks, in architectures resonating with classical neuroscientific models of hierarchical control (Merel et al., 2018; Heess et al., 2016).

Intriguingly, hierarchical deep RL systems have in some cases been configured to operate at different timescales at different levels, with slower updates at higher levels, an organizational principle that resonates with some neuroscientific evidence concerning hierarchically organized timescales across cortex (Badre, 2008; Hasson et al., 2008).

Social Cognition

A growing field of neuroscience research investigates the neural underpinnings of social cognition. In the past couple of years deep RL has entered this space, developing methods to train multiple agents in parallel in interesting multi-agent scenarios. This includes competitive team games, where individual agents must learn how to coordinate their actions (Jaderberg et al., 2019; Berner et al., 2019); cooperative games requiring difficult coordination (Foerster et al., 2019); as well as thorny "social dilemmas," where short-sighted selfish actions must be weighed against cooperative behavior (Leibo et al., 2017). The behavioral sciences have long studied such situations, and multi-agent deep RL offers new computational leverage on this area of research, up to and including the neural mechanisms underlying mental models of others, or "theory of mind" (Rabinowitz et al., 2018; Tacchetti et al., 2018).

Challenges and Caveats

It is important to note that deep RL is an active, and indeed quite new, area of research, and there are many aspects of animal and especially human behavior that it does not yet successfully capture. Arguably, from a neuroscience perspective, these limitations have an upside, in that they throw into relief those cognitive capacities that remain most in need of computational elucidation (Lake et al., 2017; Zador, 2019) and indeed point to particular places where neuroscience might be able to benefit AI research.

One issue that has already been frequently pointed out is the slowness of learning in deep RL, that is, its demand for large amounts of data. DQN, for example, required much more experience to reach human-level performance in Atari games than would be required by an actual human learner (Lake et al., 2017). This issue is more complicated than it sounds at first, both because standard deep RL algorithms have become progressively more sample efficient, through alternative approaches such as meta-learning and deep RL based on episodic memory (Ritter et al., 2018; Botvinick et al., 2019), and because human learners bring to bear a lifetime of prior experiences to each new learning problem.

Having said this, it is also important to acknowledge that deep RL systems have not yet been proved to be capable of matching humans when it comes to flexible adaptation on the basis of structured inference, leveraging a powerful store of background knowledge. Whether deep RL systems can close this gap is an open and exciting question. Some recent work suggests that deep RL systems can, under the right circumstances, capitalize on past learning to quickly adapt systematically to new situations that appear quite novel (Hill et al., 2019), but this does not invariably happen (see, e.g., Lake and Baroni, 2017), and understanding the difference is of interest both to AI and neuroscience.

A second set of issues centers on more nuts-and-bolts aspects of how learning occurs. One important challenge, in this regard, is long-term temporal credit assignment, that is, updating behavior on the basis of rewards that may not accrue until a substantial time after the actions that were responsible for generating them. This remains a challenge for deep RL systems. Novel algorithms have recently been proposed (see, e.g., Hung et al., 2019), but the problem is far from solved, and a dialog with neuroscience in this area may be beneficial to both fields.

More fundamental is the learning algorithm almost universally used in deep RL research: backpropagation. As has been widely discussed in connection with supervised deep learning research, which also uses backpropagation, there are outstanding questions about how backpropagation might be implemented in biological neural systems, if indeed it is at all (Lillicrap et al., 2020; Whittington and Bogacz, 2019; although see Sacramento et al., 2018, and Payeur et al., 2020, for interesting proposals for how backpropagation might be implemented in biological circuits). And there are inherent difficulties within backpropagation associated with preserving the results of old learning in the face of new learning, a problem for which remedies are being actively researched, in some cases taking inspiration from neuroscience (Kirkpatrick et al., 2017).

Finally, although we have stressed alignment of deep RL research with neuroscience, it is also important to highlight an important dimension of mismatch. The vast majority of contemporary deep RL research is being conducted in an engineering context, rather than as part of an effort to model brain function. As a consequence, many techniques used in deep RL research are fundamentally unlike anything that could reasonably be implemented in a biological system. At the same time, many concerns that are central in neuroscience, for example, energy efficiency or the heritability of acquired knowledge across generations, do not arise as natural questions in AI-oriented deep RL research. Of course, even when there are important aspects that differentiate engineering-oriented deep RL systems from biological systems, there may still be high-level insights that can span the divide. Nevertheless, in scoping out the potential for exchange between neuroscience and contemporary deep RL research, it is important to keep these potential sources of discrepancy in mind.

Conclusion

The recent explosion of progress in AI offers exciting new opportunities for neuroscience on many fronts. In discussing deep RL, we have focused on one particularly novel area of AI research that, in our view, has particularly rich implications for neuroscience, most of which have not yet been deeply explored. As we have described, deep RL provides an agent-based framework for studying the way reward shapes representation, and how representation in turn shapes learning and decision making, two issues that together span a large swath of what is most central to neuroscience. We look forward to an increasing engagement in neuroscience with deep RL research. As this occurs there is also a further opportunity. We have focused on how deep RL can help neuroscience, but as should be clear from much of what we have written, deep RL is a work in progress. In this sense there is also the opportunity for neuroscience research to influence deep RL, continuing the synergistic "virtuous circle" that has connected neuroscience and AI for decades (Hassabis et al., 2017).

ACKNOWLEDGMENTS

The authors were funded by DeepMind.

REFERENCES

Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al. (2019). Solving Rubik's cube with a robot hand. arXiv, arXiv:1910.07113 https://arxiv.org/abs/1910.07113.

Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/download/14858/14328.

Badia, A.P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., and Blundell, C. (2020). Never give up: Learning directed exploration strategies. In International Conference on Learning Representations. https://openreview.net/pdf?id=Sye57xStvB.

Badre, D. (2008). Cognitive control, hierarchy, and the rostro-caudal organization of the frontal lobes. Trends Cogn. Sci. *12*, 193–200.

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

Balleine, B.W., and O'Doherty, J.P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology *35*, 48–69.

Banino, A., Barry, C., Uria, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., et al. (2018). Vector-based navigation using grid-like representations in artificial agents. Nature *557*, 429–433.

Banino, A., Badia, A.P., Köster, R., Chadwick, M.J., Zambaldi, V., Hassabis, D., Barry, C., Botvinick, M., Kumaran, D., and Blundell, C. (2020). MEMO: a deep network for flexible combination of episodic memories. In International Conference on Learning Representations. https://iclr.cc/virtual_2020/poster_rJxlc0EtDr.html.

Barreto, A., Dabney, W., Munos, R., Hunt, J.J., Schaul, T., van Hasselt, H.P., and Silver, D. (2017). Successor features for transfer in reinforcement learning. In Advances in Neural Information Processing Systems, pp. 4055–4065. https://papers.nips.cc/paper/6994-successor-features-for-transfer-in-reinforcement-learning.pdf.

Barreto, A., Borsa, D., Hou, S., Comanici, G., Aygün, E., Hamel, P., Toyama, D., Mourad, S., Silver, D., Precup, D., et al. (2019). The option keyboard: combining skills in reinforcement learning. In Advances in Neural Information Processing Systems, pp. 13031–13041. https://papers.nips.cc/paper/9463-the-option-keyboard-combining-skills-in-reinforcement-learning.pdf.

Battaglia, P.W., Hamrick, J.B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., et al. (2018). Relational inductive biases, deep learning, and graph networks. arXiv, arXiv:1806.01261 https://arxiv.org/abs/1806.01261.

Behrens, T.E.J., Muller, T.H., Whittington, J.C.R., Mark, S., Baram, A.B., Stachenfeld, K.L., and Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. Neuron *100*, 490–509.

Bellemare, M.G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: an evaluation platform for general agents. J. Artif. Intell. Res. *47*, 253–279.

Bellemare, M., Srinivasan, S., Ostrovski, G., Schaul, T., Saxton, D., and Munos, R. (2016). Unifying countbased exploration and intrinsic motivation. In Advances in Neural Information Processing Systems, pp. 1471–1479. https://papers.nips.cc/paper/6383-unifying-count-based-exploration-and-intrinsic-motivation.

Bellemare, M.G., Dabney, W., and Munos, R. (2017). A distributional perspective on reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning-Volume, *70*, pp. 449–458. http://proceedings.mlr.press/v70/bellemare17a/bellemare17a.pdf.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al. (2019). Dota 2 with large scale deep reinforcement learning. arXiv, arXiv:1912.06680 https://arxiv.org/abs/1912.06680.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J.Z., Rae, J., Wierstra, D., and Hassabis, D. (2016). Model-free episodic control. arXiv, arXiv:1606.04460 https://arxiv.org/abs/1606.04460.

Bornstein, A.M., and Norman, K.A. (2017). Reinstated episodic context guides sampling-based decisions for reward. Nat. Neurosci. *20*, 997–1003.

Botvinick, M.M., and Cohen, J.D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. Cogn. Sci. *38*, 1249–1285.

Botvinick, M.M., Niv, Y., and Barto, A.G. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition *113*, 262–280.

Botvinick, M., Weinstein, A., Solway, A., and Barto, A. (2015). Reinforcement learning, efficient coding, and the statistics of natural tasks. Curr. Opin. Behav. Sci. *5*, 71–77.

Botvinick, M., Ritter, S., Wang, J.X., Kurth-Nelson, Z., Blundell, C., and Hassabis, D. (2019). Reinforcement learning, fast and slow. Trends Cogn. Sci. *23*, 408–422.

Bromberg-Martin, E.S., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. J. Neurophysiol. *104*, 1068–1076.

Burda, Y., Edwards, H., Storkey, A., and Klimov, O. (2019). Exploration by random network distillation. In International Conference on Learning Representations. https://openreview.net/pdf?id=H1lJJnR5Ym.

Carey, A.A., Tanaka, Y., and van der Meer, M.A.A. (2019). Reward revaluation biases hippocampal replay content away from the preferred outcome. Nat. Neurosci. *22*, 1450–1459.

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., and Olah, C. (2019). Exploring neural networks with activation atlases. Distill. https://distill.pub/2019/activation-atlas.

Chatham, C.H., and Badre, D. (2015). Multiple gates on working memory. Curr. Opin. Behav. Sci. *1*, 23–31.

Chentanez, N., Barto, A.G., and Singh, S.P. (2005). Intrinsically motivated reinforcement learning. In Advances in Neural Information Processing Systems, pp. 1281–1288. https://papers.nips.cc/paper/2552-intrinsically-motivated-reinforcement-learning.pdf.

Cichy, R.M., and Kaiser, D. (2019). Deep neural networks as scientific models. Trends Cogn. Sci. *23*, 305–317.

Cobbe, K., Klimov, O., Hesse, C., Kim, T., and Schulman, J. (2019). Quantifying generalization in reinforcement learning. In International Conference on Machine Learning, pp. 1282–1289. http://proceedings.mlr.press/v97/cobbe19a/cobbe19a.pdf.

Collins, A.G., and Frank, M.J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. Eur. J. Neurosci. *35*, 1024–1035.

Constantinescu, A.O., O'Reilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. Science *352*, 1464–1468.

Costa, V.D., Mitz, A.R., and Averbeck, B.B. (2019). Subcortical substrates of explore-exploit decisions in primates. Neuron *103*, 533–545.e5.

Cushman, F., and Morris, A. (2015). Habitual control of goal selection in humans. Proc. Natl. Acad. Sci. U S A *112*, 13817–13822.

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., and Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. Nature *577*, 671–675.

Dasgupta, I., Wang, J., Chiappa, S., Mitrovic, J., Ortega, P., Raposo, D., Hughes, E., Battaglia, P., Botvinick, M., and Kurth-Nelson, Z. (2019). Causal reasoning from meta-reinforcement learning. arXiv, arXiv:1901.08162 https://arxiv.org/abs/1901.08162.

Daw, N.D., and O'Doherty, J.P. (2014). Multiple systems for value learning. In Neuroeconomics, P.W. Glimcher and E. Fehr, eds. (Elsevier), pp. 393–410.

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P., and Dolan, R.J. (2011). Model-based influences on humans' choices and striatal prediction errors. Neuron *69*, 1204–1215.

Dayan, P., and Balleine, B.W. (2002). Reward, motivation, and reinforcement learning. Neuron *36*, 285–298.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 248–255.

Denil, M., Agrawal, P., Kulkarni, T.D., Erez, T., Battaglia, P., and de Freitas, N. (2016). Learning to perform physics experiments via deep reinforcement learning. arXiv, arXiv:1611.01843 https://arxiv.org/abs/1611.01843.

Diuk, C., Cohen, A., and Littman, M.L. (2008). An object-oriented representation for efficient reinforcement learning. In Proceedings of the 25th International Conference on Machine Learning, pp. 240–247. https://dl.acm.org/doi/10.1145/1390156.1390187.

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

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., and Tanila, H. (1999). The hippocampus, memory, and place cells: is it spatial memory or a memory space? Neuron *23*, 209–226.

Foerster, J., Song, F., Hughes, E., Burch, N., Dunning, I., Whiteson, S., Botvinick, M., and Bowling, M. (2019). Bayesian action decoder for deep multi-agent reinforcement learning. In International Conference on Machine Learning, pp. 1942–1951. http://proceedings.mlr.press/v97/foerster19a/foerster19a.pdf.

Frank, M.J., and Claus, E.D. (2006). Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol. Rev. *113*, 300–326.

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science *291*, 312–316.

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. *36*, 193–202.

Gelada, C., Kumar, S., Buckman, J., Nachum, O., and Bellemare, M.G. (2019). DeepMDP: learning continuous latent space models for representation learning. In International Conference on Machine Learning, pp. 2170–2179. http://proceedings.mlr.press/v97/gelada19a/gelada19a.pdf.

Gershman, S.J. (2018). Deconstructing the human algorithms for exploration. Cognition *173*, 34–42.

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.

Gershman, S.J., Blei, D.M., and Niv, Y. (2010). Context, learning, and extinction. Psychol. Rev. *117*, 197–209.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron *66*, 585–595.

Glimcher, P.W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc. Natl. Acad. Sci. U S A *108* (*Suppl 3*), 15647–15654.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). Deep Learning, *Vol. 1* (MIT Press).

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. Nature *538*, 471–476.

Güçlü, U., and van Gerven, M.A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. Front. Comput. Neurosci. *11*, 7.

Guez, A., Mirza, M., Gregor, K., Kabra, R., Racanière, S., Weber, T., Raposo, D., Santoro, A., Orseau, L., Eccles, T., et al. (2019). An investigation of model-free planning. arXiv, arXiv:1901.03559 https://arxiv.org/abs/1901.03559.

Gupta, A.S., van der Meer, M.A., Touretzky, D.S., and Redish, A.D. (2010). Hippocampal replay is not a simple function of experience. Neuron *65*, 695–705.

Ha, D., and Schmidhuber, J. (2018). World models. arXiv, arXiv:1803.10122 https://arxiv.org/abs/1803.10122.

Hamrick, J.B., Ballard, A.J., Pascanu, R., Vinyals, O., Heess, N., and Battaglia, P.W. (2017). Metacontrol for adaptive imagination-based optimization. arXiv, arXiv:1705.02670 https://arxiv.org/abs/1705.02670.

Hansen, S., Dabney, W., Barreto, A., Warde-Farley, D., de Wiele, T.V., and Mnih, V. (2020). Fast task inference with variational intrinsic successor features. In International Conference on Learning Representations https://openreview.net/pdf?id=BJeAHkrYDS.

Harb, J., Bacon, P.-L., Klissarov, M., and Precup, D. (2018). When waiting is not an option: learning options with a deliberation cost. In Thirty-Second AAAI Conference on Artificial Intelligence https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewFile/17421/16660.

Harutyunyan, A., Dabney, W., Borsa, D., Heess, N., Munos, R., and Precup, D. (2019). The termination critic. In The 22nd International Conference on Artificial Intelligence and Statistics, pp. 2231–2240. http://proceedings.mlr.press/v89/harutyunyan19a/harutyunyan19a.pdf.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. Neuron *95*, 245–258.

Hasson, U., Yang, E., Vallines, I., Heeger, D.J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. J. Neurosci. *28*, 2539–2550.

Hasson, U., Nastase, S.A., and Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. Neuron *105*, 416–434.

Hebb, D.O. (1949). The Organization of Behavior: A Neuropsychological Theory. (John Wiley).

Heess, N., Wayne, G., Tassa, Y., Lillicrap, T., Riedmiller, M., and Silver, D. (2016). Learning and transfer of modulated locomotor controllers. arXiv, arXiv:1610.05182 https://arxiv.org/abs/1610.05182.

Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., and Lerchner, A. (2017). Darla: improving zero-shot transfer in reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, *70*, pp. 1480–1490. http://proceedings.mlr.press/v70/higgins17a/higgins17a.pdf.

Hill, F., Lampinen, A., Schneider, R., Clark, S., Botvinick, M., McClelland, J.L., and Santoro, A. (2019). Emergent systematic generalization in a situated agent. arXiv, arXiv:1910.00571 https://arxiv.org/abs/1910.00571.

Hubel, D.H., and Wiesel, T.N. (1959). Receptive fields of single neurones in the cat's striate cortex. J. Physiol. *148*, 574–591.

Hung, C.-C., Lillicrap, T., Abramson, J., Wu, Y., Mirza, M., Carnevale, F., Ahuja, A., and Wayne, G. (2019). Optimizing agent behavior over long time scales by transporting value. Nat. Commun. *10*, 5223.

Jaderberg, M., Mnih, V., Czarnecki, W.M., Schaul, T., Leibo, J.Z., Silver, D., and Kavukcuoglu, K. (2016). Reinforcement learning with unsupervised auxiliary tasks. arXiv, arXiv:1611.05397 https://arxiv.org/abs/1611.05397.

Jaderberg, M., Czarnecki, W.M., Dunning, I., Marris, L., Lever, G., Castañeda, A.G., Beattie, C., Rabinowitz, N.C., Morcos, A.S., Ruderman, A., et al. (2019). Human-level performance in 3D multiplayer games with population-based reinforcement learning. Science *364*, 859–865.

Jinnai, Y., Park, J.W., Machado, M.C., and Konidaris, G. (2020). Exploration in reinforcement learning with deep covering options. In International Conference on Learning Representations https://openreview.net/pdf?id=SkeIyaVtwB.

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S.V., and McDermott, J.H. (2018). A taskoptimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron *98*, 630–644.e16.

Keramati, M., Smittenaar, P., Dolan, R.J., and Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal-directed spectrum. Proc. Natl. Acad. Sci. U S A *113*, 12868–12873.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proc. Natl. Acad. Sci. U S A *114*, 3521–3526.

Kohonen, T. (2012). Self-Organization and Associative Memory, *Vol. 8* (New York: Springer Science & Business Media).

Konidaris, G. (2019). On the necessity of abstraction. Curr. Opin. Behav. Sci. *29*, 1–7.

Konidaris, G., Osentoski, S., and Thomas, P. (2011). Value function approximation in reinforcement learning using the Fourier basis. In Twenty-Fifth AAAI Conference on Artificial Intelligence https://dl.acm.org/doi/10.5555/2900423.2900483.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. *1*, 417–446.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pp. 1097–1105. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

Kulkarni, T.D., Saeedi, A., Gautam, S., and Gershman, S.J. (2016). Deep successor reinforcement learning. arXiv, arXiv:1606.02396 https://arxiv.org/abs/1606.02396.

Kumaran, D., Hassabis, D., and McClelland, J.L. (2016). What learning systems do intelligent agents need? complementary learning systems theory updated. Trends Cogn. Sci. *20*, 512–534.

Lake, B.M., and Baroni, M. (2017). Generalization without systematicity: on the compositional skills of sequence-to-sequence recurrent networks. arXiv, arXiv:1711.00350 https://arxiv.org/abs/1711.00350.

Lake, B.M., Ullman, T.D., Tenenbaum, J.B., and Gershman, S.J. (2017). Building machines that learn and think like people. Behav. Brain Sci. *40*, e253.

Lee, D., Seo, H., and Jung, M.W. (2012). Neural basis of reinforcement learning and decision making. Annu. Rev. Neurosci. *35*, 287–308.

Lee, S.W., Shimojo, S., and O'Doherty, J.P. (2014). Neural computations underlying arbitration between model-based and model-free learning. Neuron *81*, 687–699.

Leibo, J., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. In AAMAS, *Volume 16* (ACM), pp. 464–473.

Lengyel, M., and Dayan, P. (2008). Hippocampal contributions to control: the third way. In Advances in Neural Information Processing Systems, pp. 889–896. https://papers.nips.cc/paper/3311-hippocampal-contributions-to-control-the-third-way.pdf.

Lillicrap, T.P., Santoro, A., Marris, L., Akerman, C.J., and Hinton, G. (2020). Backpropagation and the brain. Nat. Rev. Neurosci. *21*, 335–346.

Lin, L.J. (1991). Programming robots using reinforcement learning and teaching. In AAAI-91 Proceedings, pp. 781–786. https://www.aaai.org/Papers/AAAI/1991/AAAI91-122.pdf.

Lyle, C., Bellemare, M.G., and Castro, P.S. (2019). A comparative analysis of expected and distributional reinforcement learning. Proc. Conf. AAAI Artif. Intell. *33*, 4504–4511.

Machado, M.C., Bellemare, M.G., and Bowling, M. (2017). A Laplacian framework for option discovery in reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, *70*, pp. 2295–2304. http://proceedings.mlr.press/v70/machado17a/machado17a.pdf.

Mahadevan, S., and Maggioni, M. (2007). Proto-value functions: a Laplacian framework for learning representation and control in markov decision processes. J. Mach. Learn. Res. *8*, 2169–2231.

Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78–84.

Marblestone, A.H., Wayne, G., and Kording, K.P. (2016). Toward an integration of deep learning and neuroscience. Front. Comput. Neurosci. *10*, 94.

Mattar, M.G., and Daw, N.D. (2018). Prioritized memory access explains planning and hippocampal replay. Nat. Neurosci. *21*, 1609–1617.

Merel, J., Ahuja, A., Pham, V., Tunyasuvunakool, S., Liu, S., Tirumala, D., Heess, N., and Wayne, G. (2018). Hierarchical visuomotor control of humanoids. arXiv, arXiv:1811.09656 https://arxiv.org/abs/1811.09656.

Merel, J., Botvinick, M., and Wayne, G. (2019). Hierarchical motor control in mammals and machines. Nat. Commun. *10*, 5489.

Mikhael, J.G., and Bogacz, R. (2016). Learning reward uncertainty in the basal ganglia. PLoS Comput. Biol. *12*, e1005062.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv, arXiv:1312.5602 https://arxiv.org/abs/1312.5602.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. Nature *518*, 529–533.

Momennejad, I. (2020). Learning structures: predictive representations, replay, and generalization. Curr. Opin. Behav. Sci. *32*, 155–166.

Nagabandi, A., Kahn, G., Fearing, R.S., and Levine, S. (2018). Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 7559–7566. https://ieeexplore.ieee.org/document/8463189.

Niekum, S., Barto, A.G., and Spector, L. (2010). Genetic programming for reward function search. IEEE Trans. Auton. Ment. Dev. *2*, 83–90.

Niv, Y. (2009). Reinforcement learning in the brain. J. Math. Psychol. *53*, 139–154.

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

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. Distill. https://distill.pub/2017/feature-visualization.

Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped DQN. In Advances in Neural Information Processing Systems, pp. 4026–4034. https://papers.nips.cc/paper/6501-deep-exploration-via-bootstrapped-dqn.

Oudeyer, P.-Y., Kaplan, F., and Hafner, V.V. (2007). Intrinsic motivation systems for autonomous mental development. IEEE Trans. Evol. Comput. *11*, 265–286.

Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature *441*, 223–226.

Pakan, J.M., Francioni, V., and Rochefort, N.L. (2018). Action and learning shape the activity of neuronal circuits in the visual cortex. Curr. Opin. Neurobiol. *52*, 88–97.

Pandarinath, C., O'Shea, D.J., Collins, J., Jozefowicz, R., Stavisky, S.D., Kao, J.C., Trautmann, E.M., Kaufman, M.T., Ryu, S.I., Hochberg, L.R., et al. (2018). Inferring single-trial neural population dynamics using sequential auto-encoders. Nat. Methods *15*, 805–815.

Parisotto, E., Song, H.F., Rae, J.W., Pascanu, R., Gulcehre, C., Jayakumar, S.M., Jaderberg, M., Kaufman, R.L., Clark, A., Noury, S., et al. (2019). Stabilizing transformers for reinforcement learning. arXiv, arXiv:1910.06764 https://arxiv.org/abs/1910.06764.

Pathak, D., Agrawal, P., Efros, A.A., and Darrell, T. (2017). Curiosity-driven exploration by selfsupervised prediction. In International Conference on Machine Learning, pp. 2778–2787. https://dl.acm.org/doi/pdf/10.5555/3305890.3305968.

Payeur, A., Guerguiev, J., Zenke, F., Richards, B., and Naud, R. (2020). Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. bioRxiv. https://doi.org/10.1101/2020.03.30.015511.

Pfeiffer, B.E., and Foster, D.J. (2013). Hippocampal place-cell sequences depict future paths to remembered goals. Nature *497*, 74–79.

Rabinowitz, N., Perbet, F., Song, F., Zhang, C., Eslami, S.A., and Botvinick, M. (2018). Machine theory of mind. In International Conference on Machine Learning, pp. 4218–4227. http://proceedings.mlr.press/v80/rabinowitz18a/rabinowitz18a.pdf.

Rajan, K., Harvey, C.D., and Tank, D.W. (2016). Recurrent network models of sequence generation and memory. Neuron *90*, 128–142.

Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. Nat. Neurosci. *22*, 1761–1770.

Ritter, S., Wang, J.X., Kurth-Nelson, Z., Jayakumar, S.M., Blundell, C., Pascanu, R., and Botvinick, M. (2018). Been there, done that: meta-learning with episodic recall. In International Conference on Machine Learning (ICML) http://proceedings.mlr.press/v80/ritter18a/ritter18a.pdf.

Roelfsema, P.R., Lamme, V.A., and Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. Nature *395*, 376–381.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1985). Learning Internal Representations by Error Propagation. Tech. Rep. (California University San Diego, La Jolla Institute for Cognitive Science).

Sacramento, J., Costa, R.P., Bengio, Y., and Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. In Advances in Neural Information Processing Systems, pp. 8721–8732. https://papers.nips.cc/paper/8089-dendritic-cortical-microcircuits-approximate-the-backpropagation-algorithm.pdf.

Schapiro, A.C., Rogers, T.T., Cordova, N.I., Turk-Browne, N.B., and Botvinick, M.M. (2013). Neural representations of events arise from temporal community structure. Nat. Neurosci. *16*, 486–492.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. (2015). Prioritized experience replay. arXiv, arXiv:1511.05952 https://arxiv.org/abs/1511.05952.

Schmidhuber, J. (1991). Curious model-building control systems. In Proceedings of the International Joint Conference on Neural Networks, pp. 1458–1463. https://ieeexplore.ieee.org/document/170605.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., et al. (2019). Mastering Atari, go, chess and shogi by planning with a learned model. arXiv, arXiv:1911.08265 https://arxiv.org/abs/1911.08265.

Schwartenbeck, P., Fitzgerald, T., Dolan, R.J., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. Front. Psychol. *4*, 710.

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T.L., Cohen, J.D., and Botvinick, M.M. (2017). Toward a rational and mechanistic account of mental effort. Annu. Rev. Neurosci. *40*, 99–124.

Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of Go with deep neural networks and tree search. Nature *529*, 484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017a). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv, arXiv:1712.01815 https://arxiv.org/abs/1712.01815.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017b). Mastering the game of Go without human knowledge. Nature *550*, 354–359.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science *362*, 1140–1144.

Singh, S., Lewis, R.L., Barto, A.G., and Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. IEEE Trans. Auton. Ment. Dev. *2*, 70–82.

Song, H.F., Yang, G.R., and Wang, X.-J. (2017). Reward-based training of recurrent neural networks for cognitive and value-based tasks. eLife *6*, e21492.

Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. Nat. Neurosci. *20*, 1643–1653.

Stalnaker, T.A., Cooch, N.K., and Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. Nat. Neurosci. *18*, 620–627.

Stalter, M., Westendorff, S., and Nieder, A. (2020). Dopamine gates visual signals in monkey prefrontal cortex neurons. Cell Rep. *30*, 164–172.e4.

Such, F.P., Madhavan, V., Liu, R., Wang, R., Castro, P.S., Li, Y., Zhi, J., Schubert, L., Bellemare, M.G., Clune, J., et al. (2019). An Atari model zoo for analyzing, visualizing, and comparing deep reinforcement learning agents. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, pp. 3260–3267. https://www.ijcai.org/Proceedings/2019/0452.pdf.

Sussillo, D., Churchland, M.M., Kaufman, M.T., and Shenoy, K.V. (2015). A neural network that finds a naturalistic solution for the production of muscle activity. Nat. Neurosci. *18*, 1025–1033.

Sutskever, I., and Hinton, G.E. (2008). Deep, narrow sigmoid belief networks are universal approximators. Neural Comput. *20*, 2629–2636.

Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning: An Introduction (Cambridge: MIT Press).

Tacchetti, A., Song, H.F., Mediano, P.A., Zambaldi, V., Rabinowitz, N.C., Graepel, T., Botvinick, M., and Battaglia, P.W. (2018). Relational forward models for multi-agent learning. arXiv, arXiv:1809.11044 https://arxiv.org/abs/1809.11044.

Teh, Y., Bapst, V., Czarnecki, W.M., Quan, J., Kirkpatrick, J., Hadsell, R., Heess, N., and Pascanu, R. (2017). Distral: robust multitask reinforcement learning. In Advances in Neural Information Processing Systems, pp. 4499–4509. https://papers.nips.cc/paper/7036-distral-robust-multitask-reinforcement-learning.pdf.

Tesauro, G. (1994). TD-Gammon, a self-teaching backgammon program, achieves master-level play. Neural Comput. *6*, 215–219.

Vértes, E., and Sahani, M. (2019). A neurally plausible model learns successor representations in partially observable environments. In Advances in Neural Information Processing Systems, pp. 13692–13702. https://papers.nips.cc/paper/9522-a-neurally-plausible-model-learns-successor-representations-in-partially-observable-environments.pdf.

Vezhnevets, A.S., Osindero, S., Schaul, T., Heess, N., Jaderberg, M., Silver, D., and Kavukcuoglu, K. (2017). FeUdal networks for hierarchical reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning, *70*, pp. 3540–3549. http://proceedings.mlr.press/v70/vezhnevets17a/vezhnevets17a.pdf.

Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., et al. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature *575*, 350–354.

Viswanathan, G.M., Buldyrev, S.V., Havlin, S., da Luz, M.G., Raposo, E.P., and Stanley, H.E. (1999). Optimizing the success of random searches. Nature *401*, 911–914.

Wang, J.X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J.Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. (2016). Learning to reinforcement learn. arXiv, arXiv:1611.05763 https://arxiv.org/abs/1611.05763.

Wang, J.X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J.Z., Hassabis, D., and Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. Nat. Neurosci. *21*, 860–868.

Watabe-Uchida, M., Eshel, N., and Uchida, N. (2017). Neural circuitry of reward prediction error. Annu. Rev. Neurosci. *40*, 373–394.

Watters, N., Matthey, L., Bosnjak, M., Burgess, C.P., and Lerchner, A. (2019). Cobra: data-efficient model-based RL through unsupervised object discovery and curiosity-driven exploration. arXiv, arXiv:1905.09275 https://arxiv.org/abs/1905.09275.

Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwinska, A., Rae, J., Mirowski, P., Leibo, J.Z., Santoro, A., et al. (2018). Unsupervised predictive memory in a goal-directed agent. arXiv, arXiv:1803.10760 https://arxiv.org/abs/1803.10760.

Weinstein, A., and Botvinick, M.M. (2017). Structure learning in motor control: A deep reinforcement learning model. arXiv, arXiv:1706.06827 https://arxiv.org/abs/1706.06827.

Werbos, P.J. (1974). Beyond regression: new tools for prediction and analysis in the behavioral sciences. Ph.D. thesis (Harvard University).

Whittington, J.C., and Bogacz, R. (2019). Theories of error back-propagation in the brain. Trends Cogn Sci. *23*, 235–250.

Whittington, J.C., Muller, T.H., Mark, S., Chen, G., Barry, C., Burgess, N., and Behrens, T.E. (2019). The Tolman-Eichenbaum machine: unifying space and relational memory through generalisation in the hippocampal formation. bioRxiv. https://doi.org/10.1101/770495.

Wilson, M.A., and McNaughton, B.L. (1994). Reactivation of hippocampal ensemble memories during sleep. Science *265*, 676–679.

Wilson, R.C., Geana, A., White, J.M., Ludvig, E.A., and Cohen, J.D. (2014a). Humans use directed and random exploration to solve the explore-exploit dilemma. J. Exp. Psychol. Gen. *143*, 2074–2081.

Wilson, R.C., Takahashi, Y.K., Schoenbaum, G., and Niv, Y. (2014b). Orbitofrontal cortex as a cognitive map of task space. Neuron *81*, 267–279.

Wimmer, G.E., and Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. Science *338*, 270–273.

Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U S A *111*, 8619–8624.

Zador, A.M. (2019). A critique of pure learning and what artificial neural networks can learn from animal brains. Nat. Commun. *10*, 3770.

Zhang, C., Vinyals, O., Munos, R., and Bengio, S. (2018). A study on overfitting in deep reinforcement learning. arXiv, arXiv:1804.06893 https://arxiv.org/abs/1804.06893.

Zheng, Z., Oh, J., and Singh, S. (2018). On learning intrinsic rewards for policy gradient methods. In Advances in Neural Information Processing Systems, pp. 4644–4654. https://papers.nips.cc/paper/7715-on-learning-intrinsic-rewards-for-policy-gradient-methods.pdf.

Zipser, D. (1991). Recurrent network model of the neural mechanism of short-term active memory. Neural Comput. *3*, 179–193.

Zipser, D., and Andersen, R.A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature *331*, 679–684.