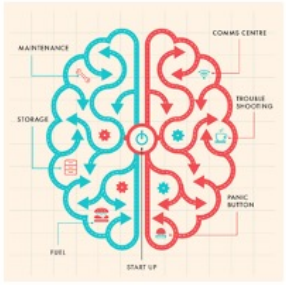
Deep reinforcement learning: Implications for neuroscience

Giuseppe di Pellegrino
Department of Psychology, University of Bologna

g.dipellegrino@unibo.it



Cognition and Neuroscience
Second cycle Degree in Artificial Intelligence – 2032/24

Deep learning and neuroscience

The past few years have seen a burst of interest in deep learning as a basis for modeling brain function, including vision, audition, motor, navigation, and cognitive control.

This interest has been catalyzed by recent dramatic advances in machine learning and artificial intelligence (AI), particularly training deep learning systems using supervised learning on tasks such as object recognition and categorization.

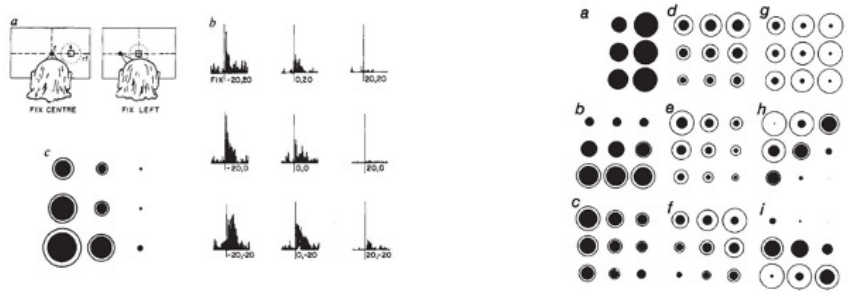
This interest of cognitive neuroscientists goes back to be the 1980s, when the first neuroscience applications of supervised deep learning began.

A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons

David Zipser* & Richard A. Andersen†

* Institute for Cognitive Science, University of California, San Diego, La Jolla, California 92093, USA
† Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Neurons in area 7a of the posterior parietal cortex of monkeys respond to both the retinal location of a visual stimulus and the position of the eyes and by combining these signals represent the spatial location of external objects. A neural network model, programmed using back-propagation learning, can decode this spatial information from area 7a neurons and accounts for their observed response properties.



Zipser & Andersen, Nature, 1988

Reinforcement Learning

RL (Sutton and Barto, 2018) considers the problem of an agent embedded in an environment, where the agent must progressively improve the actions it selects in response to each environmental situation or state.

Critically, in contrast to supervised learning, the agent does not receive explicit feedback directly indicating correct actions.

Early work on RL involved simple environments comprising few possible states and 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.

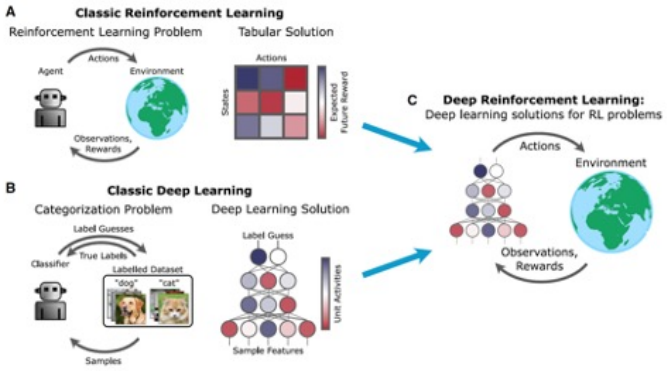
One important approach to attaining generalization across states is referred to as function approximation, which attempts to assign similar representations to states in which similar actions are required.

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.

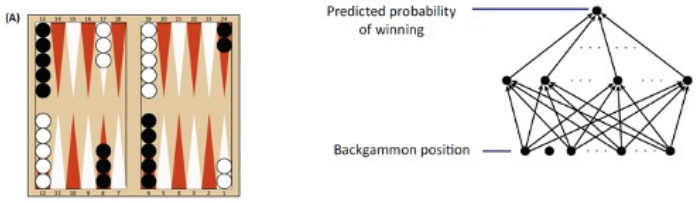
A long-standing aspiration of RL has been to perform adaptive non-linear function approximation using deep neural networks.

Deep Reinforcement Learning

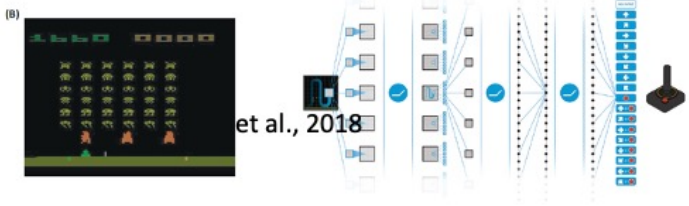
Deep RL refers to a system that solves RL problems (i.e., maximizes long-term reward), using representations that are themselves learned by a deep neural network (rather than stipulated by the designer).



Deep RL, 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.

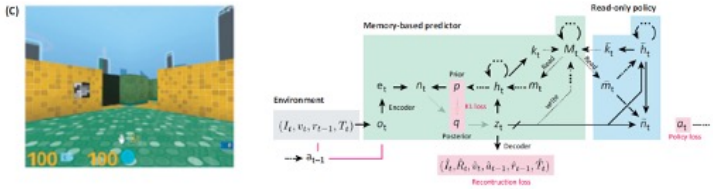


Tesauro, 1994



et al., 2018

Mnih et al., 2015

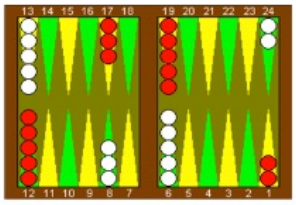


Wayne et al., 2018

Representative examples of DRL

TD-Gammon, A Self-Teaching Backgammon Program, Achieves Master-Level Play

Gerald Tesauro
IBM Thomas J. Watson Research Center
P. O. Box 704
Yorktown Heights, NY 10598
(tesauro@watson.ibm.com)



TD-Gammon combined neural networks with RL to learn from scratch how to play backgammon competitively with top human players

TD-Gammon provided a powerful example of what RL implemented via neural networks might deliver, its approach however 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.

Tesauro, Neural Comput., 1994

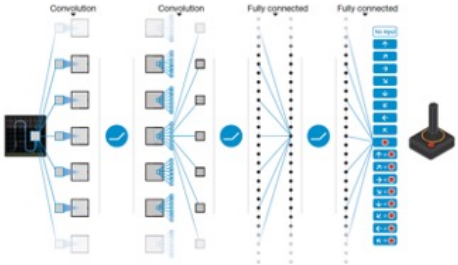
LETTER

doi:10.1038/nature14236

Human-level control through deep reinforcement learning

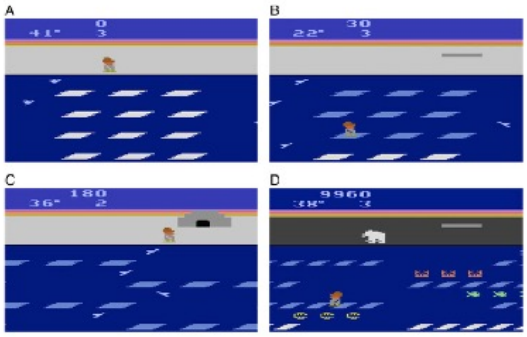
Volodymyr Mnih[1]*, Koray Kavukcuoglu[1]*, David Silver[1]*, Andrei A. Rusu[1], Joel Veness[1], Marc G. Bellemare[1], Alex Graves[1], Martin Riedmiller[1], Andreas K. Fidjeland[1], Georg Ostrovski[1], Stig Petersen[1], Charles Beattie[1], Amir Sadik[1], Ioannis Antonoglou[1], Helen King[1], Dharshan Kumaran[1], Daan Wierstra[1], Shane Legg[1] & Demis Hassabis[1]

This state of affairs changed dramatically 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)



Mnih et al., Nature, 2015

Although DQN learns to play games at human-level performance while assuming very little prior knowledge, the DQN may be learning to play Frostbite and other games in a very different way than people do.



One way to examine the differences is by considering the amount of experience required for learning.

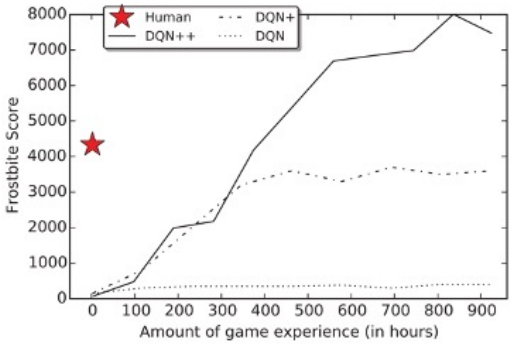
In Mnih et al. (2015), the DQN was compared with a professional gamer who received approximately 2 hours of practice on each of the 49 Atari games.

The DQN was trained on 200 million frames from each of the games, which equates to approximately 924 hours of game time (about 38 days), or almost 500 times as much experience as the human received. Additionally, the DQN incorporates experience replay, where each of these frames is replayed approximately eight more times on average over the course of learning.

More recent variants of the DQN perform better and can even outperform humans reaching 172% by using smarter replay and more efficient parameter sharing (Wang et al., 2016) but they require a lot of experience to reach this level.

The learning curve for the model of Wang et al. (2016) shows performance is approximately 44% after 200 hours, 8% after 100 hours, and less than 2% after 5 hours (which is close to random play, approximately 1.5%).

Lake et al., Brain & Behav. Sci., 2017



Learning speed for people vs. DQNs e on the Atari 2600 game Frostbite is plotted as a function of game experience (hours).

It has been argued that these machines should:

(1) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems;

(2) ground learning in intuitive theories of physics and psychology to support and enrich the knowledge that is learned; and

(3) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations.

Lake et al., Brain & Behav. Sci., 2017

Deep RL models are sample-inefficient, requiring large amounts of data to learn

Deep RL systems learn in a fashion quite different from humans.

The hallmark of this difference lies in the sample efficiency of human learning versus deep RL, that is the amount of data required for a learning system to attain any chosen target level of performance.

There are two factor that may be responsible of the sample efficiency problem of early Deep RL models: incremental parameter adjustment and weak inductive bias.

Sources of slowness in Deep RL: 1

Let' s examine some techniques that enable fast deep RL and consider their potential implications for psychology and neuroscience

One key source of slowness in deep RL is the requirement for incremental parameter adjustment - based on gradient descent - to modify the connectivity of a DNN mapping from perceptual inputs to action outputs.

The adjustments must be of small step-sizes in order to maximize generalization and avoid overwriting the effects of earlier learning (an effect sometimes referred to as 'catastrophic interference', McCloskey & Cohen, 1989).

Sources of slowness in Deep RL: 2

Learning systems necessarily faces a bias–variance trade-off:
the stronger the initial assumptions the learning procedure makes about the patterns to be learned (i.e., the stronger the initial inductive.

A learning procedure with weak inductive bias will be able to master a wider range of patterns (greater variance) but will in general be less sample-efficient.

A learning system that only considers a narrow range of hypotheses when interpreting incoming data will hone in on the correct hypothesis more rapidly than a system with weaker inductive biases.

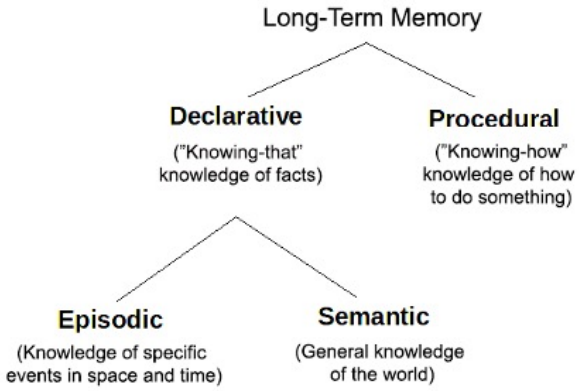
Episodic Deep RL

Memory processing might help models of RL to escape some of their previous weaknesses and inefficiencies under which they operate.

This idea, referred to as episodic RL, parallels 'non-parametric' approaches in machine learning, and 'instance-' or 'exemplar-based' theories of learning in psychology.

The agent keeps an explicit record of past events and use this record directly as a point of reference in making new decisions.

In novel and uncertain situations (e.g., complex state spaces, with very little data) the procedure is to compare a representation of the current situation with stored representations of previous situations.



Cognitive neuroscience reveals multiple memory systems, each of which possesses several properties that are relevant for decision-making.

Multiple memory system in the brain

Complementary Learning Systems (CLS) theory (McClelland, J.L. et al., 1995) holds that intelligent agents must possess two learning systems, instantiated in neocortex and hippocampus

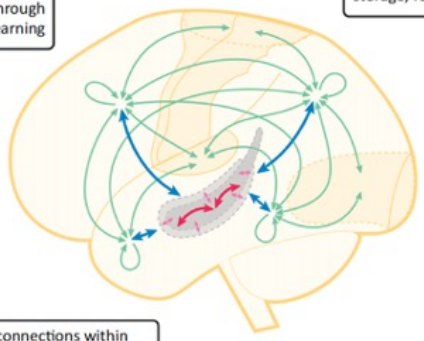
The first system allows the gradual acquisition of structured knowledge about the environment, stored in the connections among the neurons in the neocortex. This knowledge is shaped by the statistics of the environment, with dense similarity-based coding, emphasizing shared structures between experiences.

The other system, centered on the hippocampus, allows rapid learning of the spatial and non-spatial aspects of specific experiences.

Complementary Learning Systems (CLS) and their Interactions.

Connections within and among neocortical areas (green) support gradual acquisition of structured knowledge through interleaved learning

Bidirectional connections (blue) link neocortical representations to the hippocampus/MTL for storage, retrieval, and replay



Rapid learning in connections within hippocampus (red) supports initial learning of arbitrary new information

Episodic control: the third way to action selection

Pioneer work of Lengyel & Dayan (2007) suggested that episodic memories could be used to record and later mimic previously rewarding sequences of states and actions, a process they called episodic control.

Pritzel et al (2019) propose Neural Episodic Control (NEC), an agent that stores each encountered state along with the discounted sum of rewards obtained during the next n time steps. These two stored items comprise an episodic memory of the encountered state and the reward that followed.

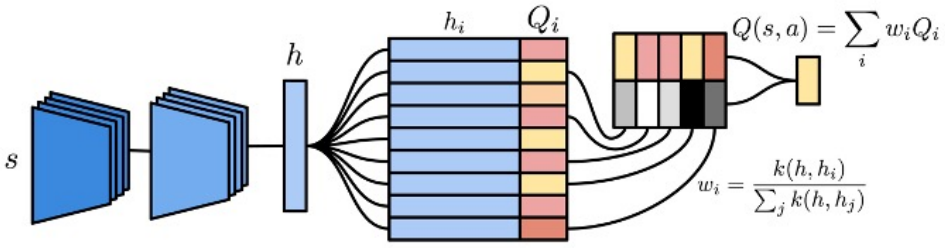
Such an episodic RL algorithm achieves strong performance on Atari games.

They employed an extremely simple model of episodic memory, and assumed that each time the subject experiences a reward that is considered large enough (larger than expected a priori)
it stores the specific sequence of state-action pairs leading up to this reward, and tries to follow such a sequence whenever it stumbles upon a state included in it.
If multiple successful sequences are available for the same state, the one that yielded maximal reward is followed.

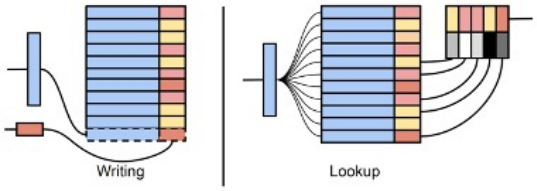
Such a strategy is expected to be useful in the low data limit because, unlike in cache-based control, there is no issue of bootstrapping and temporal credit assignment, and unlike in model-based control, there is
no exhaustive tree-search involved in action selection.

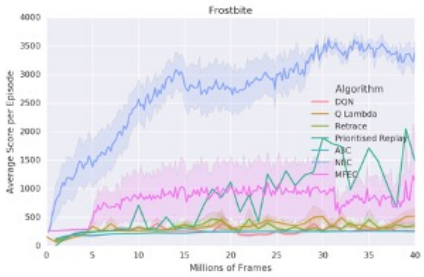
NEC consists of three components: a convolutional neural network that processes pixel images s, a set of memory modules (one per action, differentiable neural dictionary or DND), and a final network that converts read-outs from the action memories into Q(s; a) values.

To estimate the value of a new state, the agent computes a sum of the stored discounted rewards, weighted by the similarity between stored states and the new state.



There are two operations possible on a DND: lookup and write





ECN performance on Frostbite

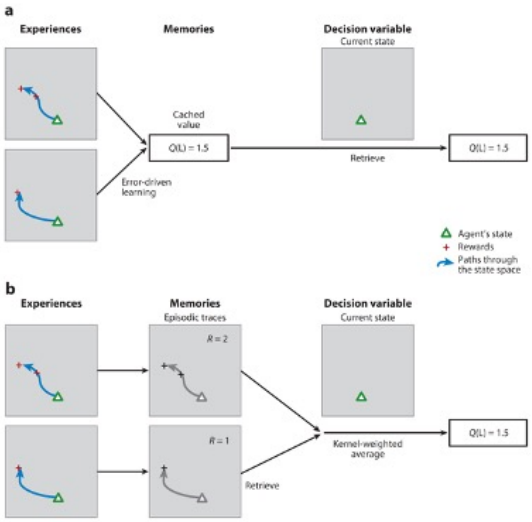
Algorithm 1 Neural Episodic Control

$\mathcal{D}$: replay memory.
M_a: a DND for each action a.
N: horizon for N-step Q estimate.
for each episode **do**
  for $t = 1, 2, \ldots, T$ **do**
    Receive observation s_t from environment with embedding h.
    Estimate $Q(s_t, a)$ for each action a via (1) from M_a
    $a_t \leftarrow \epsilon$-greedy policy based on $Q(s_t, a)$
    Take action a_t, receive reward r_{t+1}
    Append $(h, Q^{(N)}(s_t, a_t))$ to M_{a_t}.
    Append $(s_t, a_t, Q^{(N)}(s_t, a_t))$ to $\mathcal{D}$.
    Train on a random minibatch from $\mathcal{D}$.
  end for
end for

Schematic representation of different approaches to value computation

(a) In model-free RL, individual experiences are integrated into a cached value and stored in memory, which is then used to compute action values in a new state.

(b) In episodic RL, individual experiences, along with their associated returns, are retained in memory and retrieved at choice time. Each episodic trace is weighted by its similarity to the current state according to a kernel function.

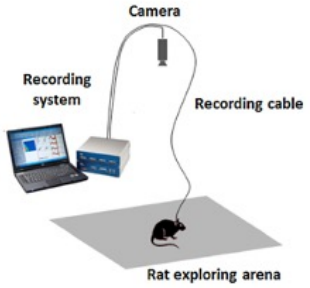


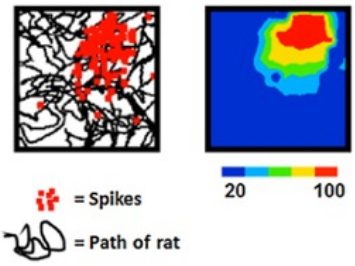
Replay

Other deep RL methods keep a history of previous experience. Indeed, DQN (Mnih et al., 2015) itself has an elementary form of memory: the replay buffer, that is frequently replayed to distil the contents into DQN's value network.

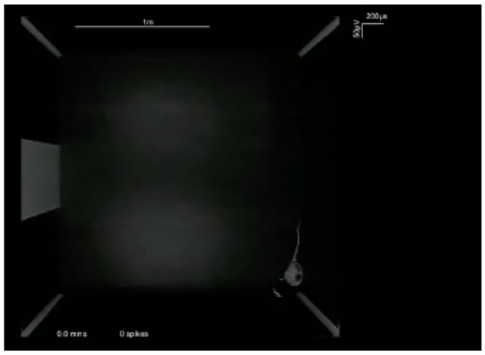
Kumaran et al. (2016) suggest that training on replayed experiences from the replay buffer in DQN is similar to the replay of experiences from episodic memory in animals.

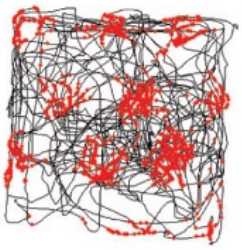
Hippocampus

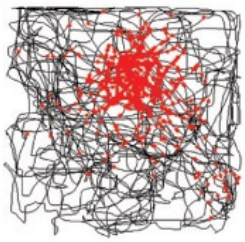




Recording of single units in the hippocampus (a deep brain structure resembling a seahorse) allowed the identification of place cells, which respond when the animal occupies specific positions in the environment (O'Keefe & Dostrovsky, 1971).

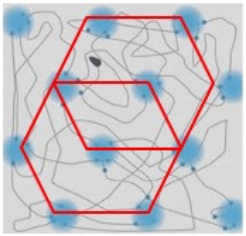


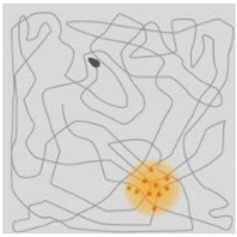




grid cell

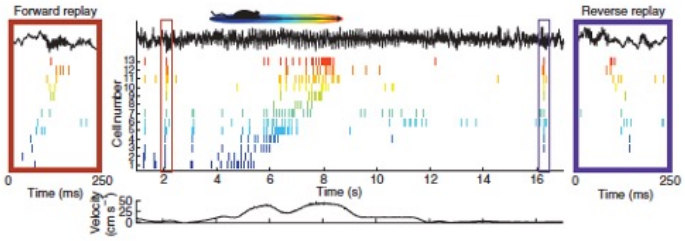
place cell





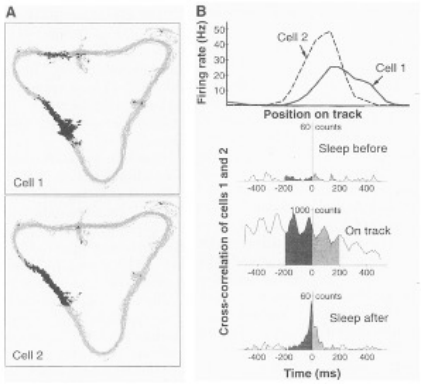
Place cells in the hippocampus and grid cells in the medial entorhinal cortex; In the figure, the receptive fields (the regions of space in which the neuron is active, in red) are superimposed on the animal's trajectory in the recording arena (black line).

While most place cells are active for a single position, grid cells have multiple zones of activation, arranged in a regular manner to form a matrix of triangular/hexagonal units that covers the entire environment available to the animal.



Replay is the sequential reactivation of hippocampal place cells that represent previously experienced behavioral trajectories and occurs frequently in sleep and wakeful rest.

The repetition of learned sequences on a compressed time scale is well suited to promote memory consolidation in distributed circuits beyond the hippocampus, suggesting that consolidation occurs in both the awake and sleeping animal.



Reminders of past choices bias decisions for reward in humans

Bornstein et al., provide evidence that decisions are made by consulting memories for individual past experiences

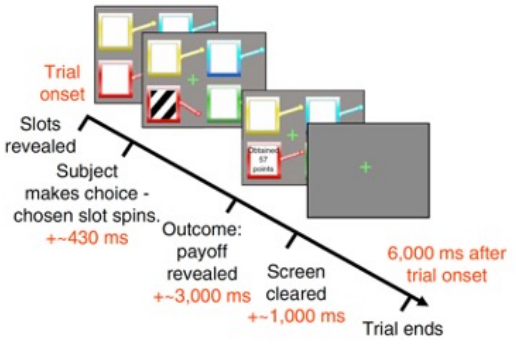
First, in a standard rewarded choice task, they show that a model that estimates value at decision-time using individual samples of past outcomes fits choices and decision related neural activity better than a canonical incremental learning model.

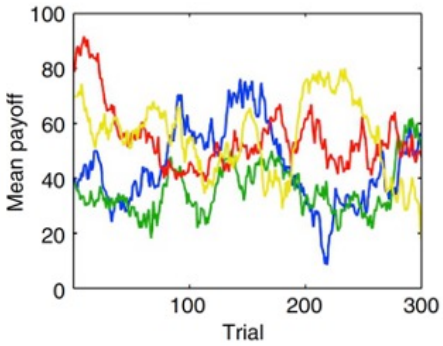
Second, they bias this sampling process by incidentally reminding participants of individual past decisions. The next decision after a reminder shows a strong influence of the action taken and value received on the reminded trial.

These results provide support for a decision architecture that relies on samples of individual past choice episodes rather than incrementally averaged rewards in evaluating options

Bornstein et al., Nat. Commun. 2017

Experiment 1





Twenty participants completed a four-choice bandit task, choosing in each of 300 trials between four different slot machines and receiving a payoff (between 0 and 100 points) for their choice

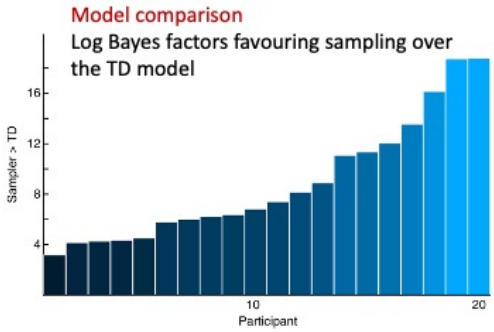
Bornstein et al., Nat. Commun. 2017

Two distinct types of models (1 and 2) were compared in their effectiveness at explaining each participant's time series of choices:

1. implemented a TD learning approach that kept a running average estimate of action value;

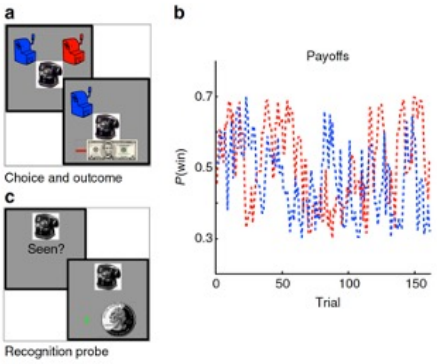
2. followed a strategy of sampling from previous experiences to estimate these values at the time of decision. In this model, the most recent experience is most likely to be sampled and previous trials are successively exponentially less likely to be sampled.

For all 20 participants individually, choices were better fit by the sampling model than by the incremental learning model



Bornstein et al., Nat. Commun. 2017

Experiment 2



In Exp 1, one cannot directly observe which individual trials participants have sampled.

To address this, a choice-incidental information was used to tag each choice as a unique event.

Interspersed among the 130 choices were 32 recognition probe trials on which participants were asked whether or not they recognized a given 'ticket' image.

These probes were intended to bring to mind the specific trial on which the probed image was experienced.

Bornstein et al., Nat. Commun. 2017

Specifically, the bandit task was modified so that each trial involved a unique photograph of an object (a "ticket released by the slot machine); the number of choice options was reduced to two, and outcome values were limited to wins or losses of $5. The probability that each machine would pay a winning ticket varied from trial to trial.

The key test of hypothesis is whether choices following a probe show an effect of the cued trial.

For example, if a given recognition memory probe evoked a trial on which the participant chose the blue bandit and was rewarded, then the participant should be more likely to choose the blue bandit on the subsequent choice trial.

Consistent with this hypothesis, choices following a memory probe were significantly influenced by experience evoked by the probed ticket.

Probes evoked choices that were, on average, 39 trials in the past—a temporal horizon minimizing the likelihood that the reminded trial would still be present in working memory

Meta-RL

A second source of slowness in standard deep RL is weak inductive bias.

Fast learning requires the learner to start with a set of hypotheses/knowledge concerning the structure of the tasks that it will face.

One possibility is to draw on past experience.

Meta-learning refers to the leveraging of past experience to accelerate new learning

The idea originates from psychology, where it has been called 'learning to learn.'

Human and animal learners gain their efficiency at least in part from the fact that they do not learn entirely 'from scratch.' Rather than operating as a tabula rasa, biological learners bring a wealth of past learning to bear on any new learning problem, and it is precisely their preexisting knowledge that enables rapid new learning. Psychologists have labeled this phenomenon learning to learn and meta-learning.

Theories of meta-learning, whether from a biological or artificial intelligence perspective, share the core idea that learning exploit relevant past experience, rather than begin anew with each new task.

Behaviorally, the signature of meta-learning is the speed-up of learning with repeated experience in a domain, this has been variously described as the formation of 'learning sets' (Harlow, 1949) or the integration of experience into structured knowledge known as a 'schema' (Bartlett, 1932, Tse et al., 2007).

Harlow (1949) presented monkeys with two unfamiliar objects, and permitted to grab one of them. Beneath lay either a food reward or an empty well.
The objects were then placed before the animal again, possibly left–right reversed, and the procedure was repeated for a total of six rounds.
Two new and unfamiliar objects were then substituted in, and another six trials ensued with these objects. Then another pair of objects, and so forth.
Across many object pairs, the animals were able to figure out that a simple rule always held: one object yielded food and the other did not, regardless of position.
When presented with a new pair of objects, Harlow's monkeys were able to learn in one shot which the preferable object was, a simple but vivid example of learning to learn.

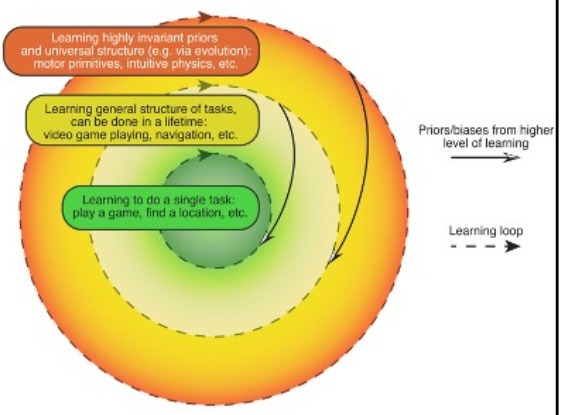
Harlow H.F., The formation of learning sets. Psychol. Rev. 1949; 56: 51

Multiple nested scales of learning in nature

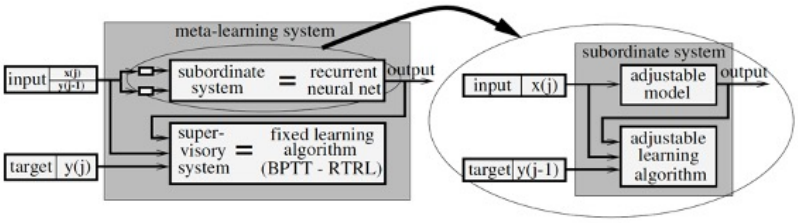
At the highest level, learning is done via evolution, to learn highly invariant universal structure such as intuitive physics, motor primitives, or other kinds of 'core knowledge'.

At the level below, learning is done within a lifetime and involves learning the general structure of different tasks, such as video game playing, how to navigate around a city, or acquiring specific skills.

Learning at the innermost level involves fast adaptation within a specific task, such as playing a new video game or finding a certain restaurant within a new city.

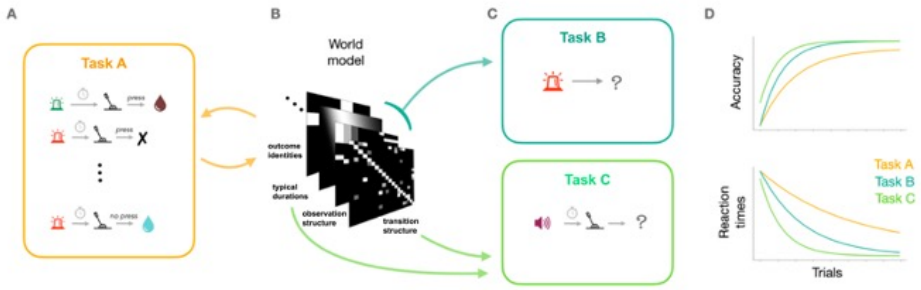


One of the first computational approach was introduced by Hochreiter and colleagues (2001), in which a recurrent neural network is trained on a series of interrelated tasks using standard backpropagation



A critical aspect of their setup is that the network receives, on each step within a task, an auxiliary input indicating the target output for the preceding step.

For example, in a regression task, on each step the network receives as input an x value for which it is desired to output the corresponding y, but the network also receives an input disclosing the target y value for the preceding step.



Model-based RL algorithms in which task representations are learned concurrently with reward expectations are naturally suited for meta-learning, in that any or all properties of the learned internal model of the tasks can be selectively generalized to a new setting

For example, building a model of the transition structure between states of the task, the different observation probabilities associated with these states, likely state durations and expectations about outcome identities can both facilitate learning in a single task environment (task A) and provide structured knowledge that can be applied to other tasks (here, task B and C).

In this way, learning in a new task may be started with relevant priors about the contingent, temporal, effortful and attentional requirements of a whole class of problems, accelerating learning and supporting rapid behavioral adaptation to a novel environment.

No special dedicated mechanism is necessary to support meta-learning.

Meta-learning emerges spontaneously from more basic mechanisms.

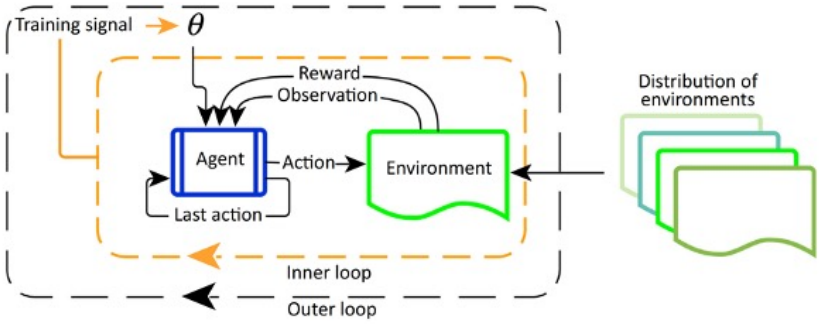
The two necessary ingredients are:
(1) a learning system that has some form of short-term memory;
(2) a training environment that exposes the learning system not to a single task, but instead to a sequence or distribution of interrelated tasks.

When these two ingredients are simultaneously present, something remarkable occurs: The system slowly learns to use its short-term memory as a basis for fast learning.

Meta-RL

Meta-learning occurs when one learning system progressively adjusts the operation of a second learning system, such that the latter operates with increasing speed and efficiency.

This is often described in terms of an outer-loop (slow) and inner-loop (fast) learning systems: the 'outer loop' (slow) uses its experiences over many task contexts to gradually adjust parameters that govern the operation of an 'inner loop' (fast), so that the inner loop can adjust rapidly to new tasks.



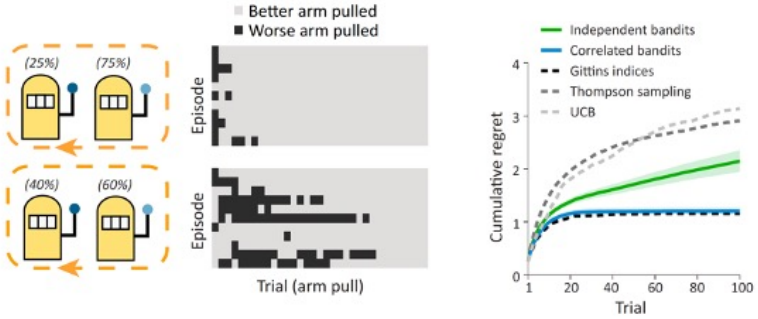
In contrast to the generality of reinforcement learning algorithms defined over broad classes of environments (i.e. all Markov Decision Processes), meta-learning allows the agent to adjust its learning process to specialize to a more narrow distribution of environments, sacrificing complete generality for rapid adaptation to a set of target environments.

The fast learning system represents the agent's learning process that occurs during interaction with a particular instance of an environment, determining the agent's online behavior.
This learning system is the one which is adapted specifically to the types of environments the agent designer is targeting for rapid adaptation.

At the opposite end of the spectrum of temporal horizons, the slow learning system represents, the

complementary part of the agent taking experiences aggregated across many different trials in order to tailor the fast learning system's inductive biases to the distribution of target environments.

Wang et al. (2016) trained a recurrent neural network on a series of interrelated RL tasks (i.e., bandit problems) varying only in their parameterization.



The agent interacts with one bandit problem for a fixed number of steps and then moves on to another. The challenge for the agent, in each new bandit problem, is to balance exploration of the two alternatives with exploitation of information gained so far, seeking to maximize cumulative reward.

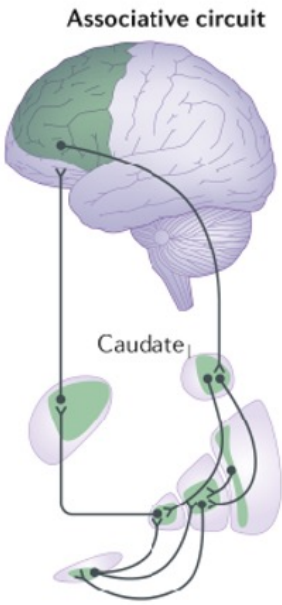
Meta-RL learns to solve this problem, learning to explore more on a difficult bandit problem with arm reward probabilities that are more closely matched [(40%, 60%) versus (25%, 75%); compare lower panel with upper panel of Figure.

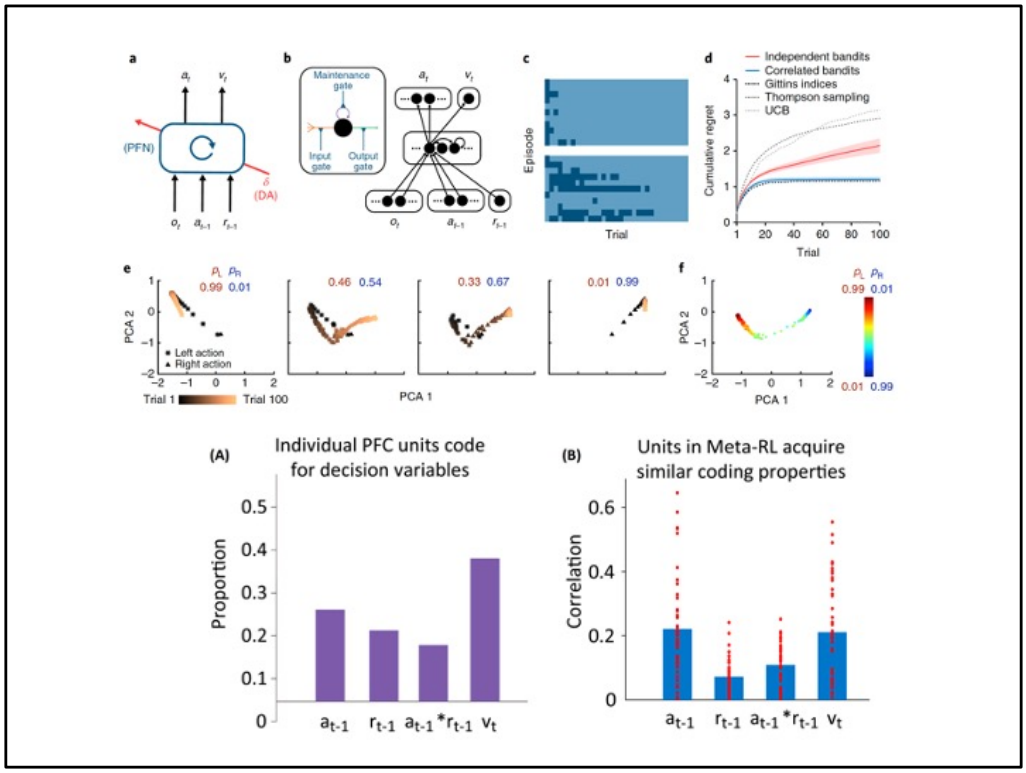
After training on a number of problems, the network can, even with its connection weights fixed, explore a new bandit problem

Wang and colleagues (2018) have proposed a very direct mapping from the elements of meta-RL to neural structures and functions.

Specifically, they propose that slow, dopamine-driven synaptic change may serve to tune the activity dynamics of prefrontal circuits, in such a way that the latter come to implement an independent set of learning procedures.

Through a set of computer simulations, Wang and colleagues demonstrated how meta-RL, interpreted in this way, can account for a diverse range of empirical findings from the behavioral and neurophysiology literatures.




a
b
c
d
(PFN)
(DA)
Maintenance gate
Input gate
Output gate
Episode
Trial
Independent bandits
Correlated bandits
Gittins indices
Thompson sampling
UCB
Cumulative regret
e
f
PCA 2
PCA 1
Left action
Right action
Trial 1
Trial 100
0.46 0.54
0.33 0.67
0.01 0.99
0.99 0.01
(A)
Individual PFC units code for decision variables
Proportion
(B)
Units in Meta-RL acquire similar coding properties
Correlation
a_{t-1} r_{t-1} $a_{t-1}*r_{t-1}$ v_t