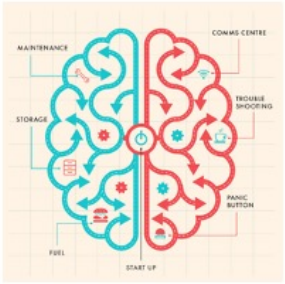
Dopamine coding of Reward Prediction Error

Giuseppe di Pellegrino
Department of Psychology, University of Bologna

g.dipellegrino@unibo.it



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

What is decision-making (DM)?

DM is a deliberative (voluntary) process that results in the selection of an action, based on sensory information (about the environment and the agent's internal state).

The study of decision-making spans several fields, including neuroscience, psychology, statistics, economics, computer science, philosophy, medicine, and jurisprudence.

The neuroscience of decision making is really the neuroscience of cognition, as it concerns principles of neural processing that underlie a variety of mental functions, relevant to both health and disease.

Decisions are inherently probabilistic not deterministic

First, agents make inconsistent choices.

Second, agents makes choices without fully knowing the consequences of those choices (decisions under uncertainty).

Third, both external and internal signals are corrupted by noise.

Many modern decision theories consider the choice process, including deliberations and estimation of competing outcomes as intrinsically random.

Within an evolutionary framework, it is possible to show that being stochastic may actually augment the chances of survival and reproductive success in the changing, uncertain dangerous environments that characterized homo sapiens' evolution over most of the last 200,000 years.

We distinguish two types of DM

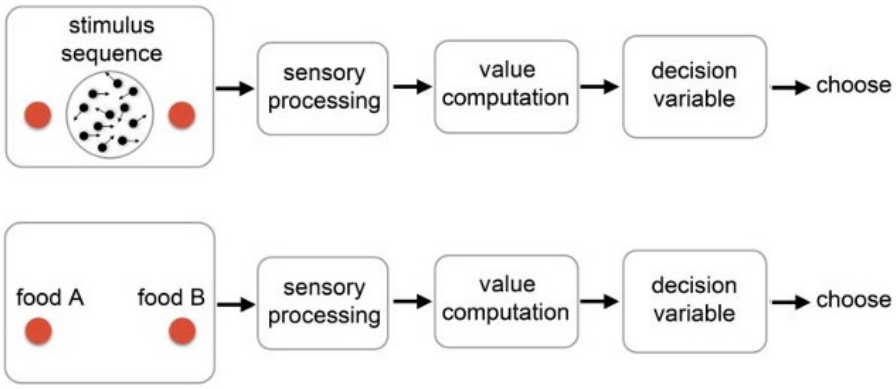
Perceptual DM: agents select action A or B based on weak or noisy external signals (do you see apple or orange?)

Value-based DM: agents select action A or B based on their (subjective) preference or value (do you prefer apple or orange?)

Source of uncertainty differs between perceptual and value-based DM

Perceptual DM: It is the identity of the stimulus that is uncertain, not the value of its associated action.

Value-based DM: It is the value of its associated action that is uncertain, not the identity of the stimulus.



We distinguish two types of DM

Perceptual DM: agents select action A or B based on weak or noisy external signals (do you see apple or orange?)

Value-based DM: agents select action A or B based on their (subjective) preference or value (do you prefer apple or orange?)

DM comprises five basic processes:

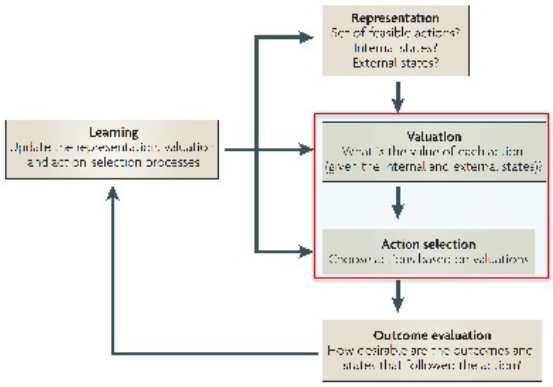
1) Representation: the available options (e.g., states) and actions, as well as internal (e.g., hunger level), and external (e.g., threat level) factors are identified;

2) Valuation: a value is assigned to the different alternatives;

3) Choice: value are compared and the action associated with highest valued is selected;

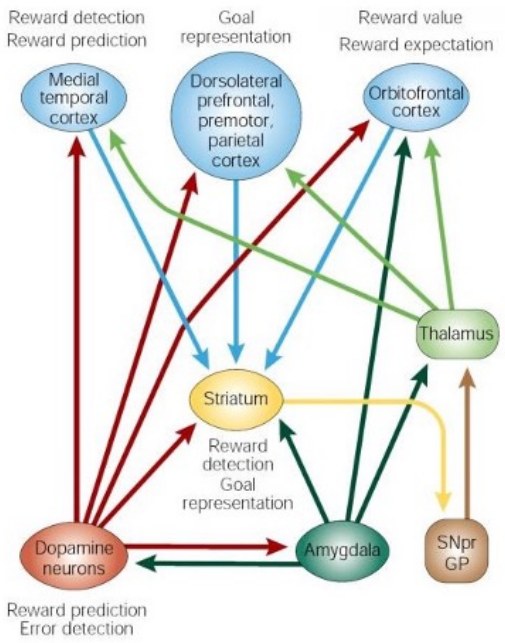
4) Outcome evaluation: after the choice, the desirability of the consumed outcomes is measured ;

5) Learning: feedback signals are used to update the relevant processes and improve the quality of future decisions;



The valuation mechanism: learns, stores, and retrieves the values of goods or actions under consideration,

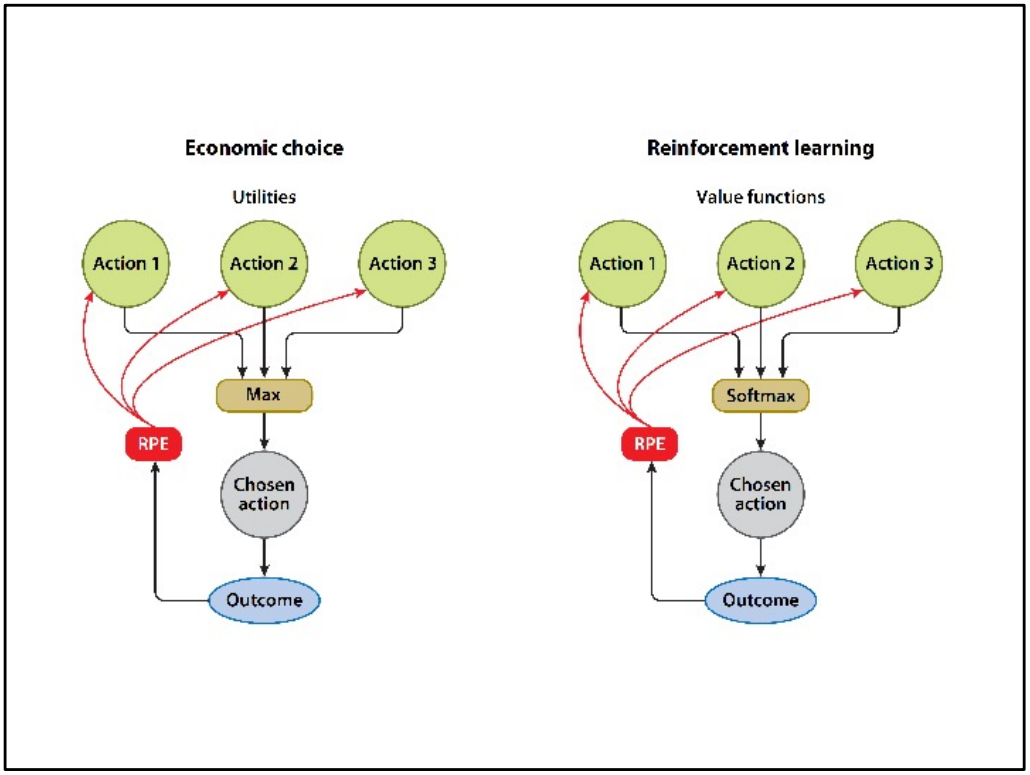
The choice mechanism: takes the output of the valuation circuit as its starting point and generates an actual choice from amongst those options.



Valuation circuitry

Neurons sensitive to reward value are widespread in the brain, including both cortical (orbitofrontal, premotor and prefrontal cortex) and subcortical (striatum, amygdala, and dopamine neurons) regions.

Schultz, Nat Reviews Neurosci, 2000



Economic and reinforcement learning theories of decision making.
In economic theories, decision-making corresponds to the selection of an action with maximum utility.
In reinforcement learning, actions are chosen probabilistically (i.e., softmax) on the basis of their value functions.
In addition, value functions are updated on the basis of the outcome (reward or penalty) resulting from the action chosen by
the animal. RPE, reward prediction error.

Reinforcement learning

We learn by interacting with our environment.

Reinforcement learning is learning what to do - how to map states to actions – in order to maximize reward.

The agent's goal is to choose actions so as to maximize the expected cumulative future rewards.

Each action not only affects the current reward but, by affecting the next state, also sets the stage for subsequent rewards, so that choosing optimally is not easy.

At any time step t, all future states and rewards depend only on the current state and action, not on all preceding events (Markov conditional independence property).

The expected future reward for taking action a_t in state s_t (following some policy π) is given by the sum of two terms, the current reward all the remaining discounted future rewards.

$$Q_\pi(s_t, a_t) = r_t + \gamma \sum_{s_{t+1}} P(s_{t+1} \mid s_t, a_t) Q_\pi(s_{t+1}, \pi(s_{t+1})).$$

There are two main classes of algorithms for RL based on Equation above; these classes focus on either the left- or right-hand side of the equal sign in that equation.

On the one hand, it is possible to iteratively expand the right-hand side of to compute the state-action value for any state and candidate action.

This requires the one-step reward and state transition distributions.

Model-Free RL

The second class of algorithms avoids learning a world model and instead learns state-action values Q directly from experience.

Such model-free RL algorithms [in particular, the family of temporal-difference (TD) learning algorithms; Sutton 1988] use experienced states, actions, and rewards to approximate the righthand side of the above Equation and average these to update a table of long-run reward predictions.

More particularly, many algorithms are based on the TD reward prediction error occasioned by comparing the value $Q(s_t, a_t)$ to a sample computed one time step later.

$$\delta_t = r_t + \gamma\, Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t).$$

There is strong evidence that the dopaminergic system is the major neural substrate associated to RL for both natural rewards and addictive drugs.

Dopamine projections include:

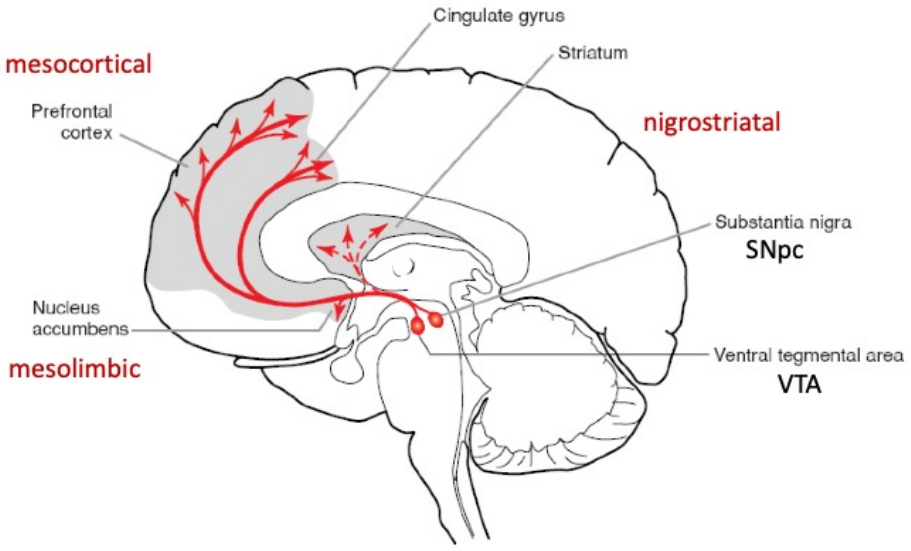
- nigrostriatal system, originates in the zona compacta of the substantia nigra (SNpc); and is identified most strongly with motor function (action policy). It projects primarily to the caudate–putamen.

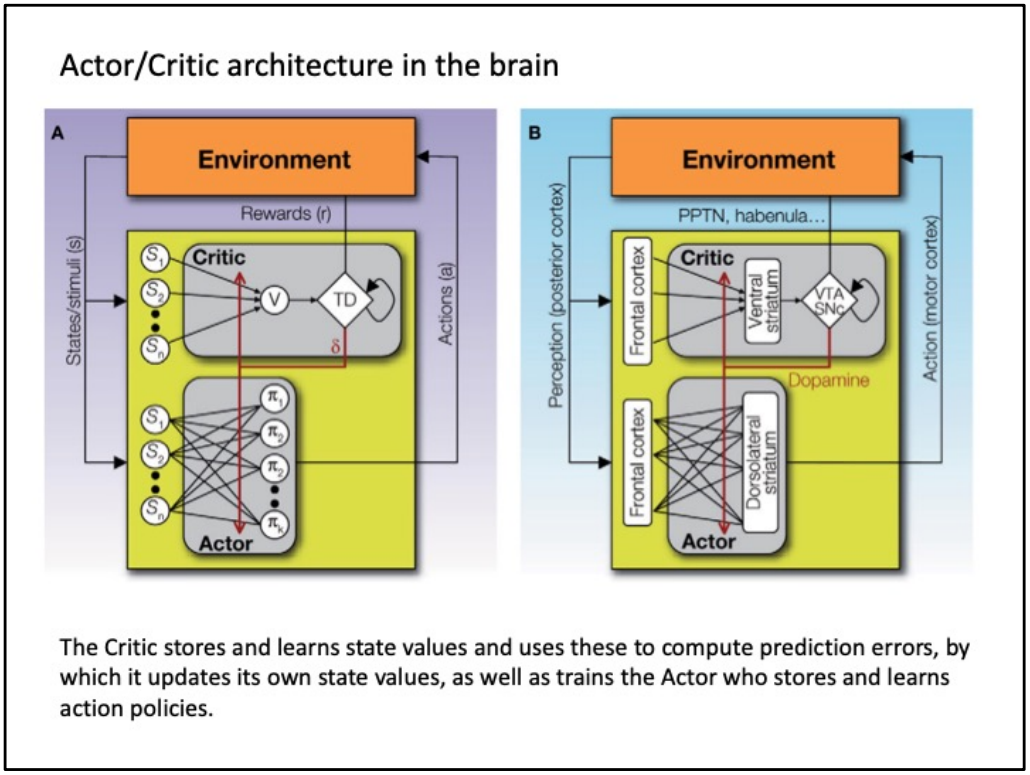
- mesocorticolimbic system, important for motivational, value function, arises from the dopamine cells in the ventral tegmental area (VTA) and comprises:

    - mesolimbic: cells in the VTA project to the nucleus accumbens, septum, amygdala and hippocampus;

    - mesocortical: cells in the VTA project to the medial prefrontal, cingulate, orbitofrontal and perirhinal cortex.

Dopaminergic system





The basic Actor/Critic architecture and its suggested neural implementation.
(A) The (external or internal) environment provides two signals to the system: S, indicating the current state or stimuli, and r indicating the current reward.
The Actor comprises of a mapping between states S and action policies $\pi(a|S)$ (through modifiable weights or associative strengths).
Its ultimate output is an action which then feeds back into the environment and serves to (possibly) earn rewards and change the state of the environment.

The Critic comprises of a mapping between states S and values V (also through modifiable weights). The value of the current state provides input to a temporal difference (TD) module that integrates the value of the current state, the value of the previous state (indicated by the feedback arrow) and the current reward, to compute a prediction error signal
$\delta t = r(St) + V(St+1) - V(St)$. This signal is used to modify the mappings in both the Actor and the Critic.
(B) A suggested mapping of the Actor/Critic architecture onto neural substrates in the cortex and basal ganglia. The mapping between states and actions in the Actor is realized through plastic synapses between the cortex and the dorsolateral striatum. The mapping between states and their values is realized though similarly modifiable synaptic strengths in cortical projections to the ventral striatum.
The prediction error is computed in the ventral tegmental area (VTA) and the substantia nigra pars compacta (SNc) – the two midbrain dopaminergic

nuclei – based on state values from ventral striatal afferents, and outcome (reward) information from sources such as the pedunculopontine nucleus (PPTN), the habenula etc.
Nigrostriatal and mesolimbic dopaminergic projections to the dorsolateral and ventral striatum, respectively, are used to modulate synaptic plasticity according to temporal difference learning.

in the Actor/Critic architecture the Critic stores and learns state values and uses these to compute prediction errors, by which it updates its own state values. These same prediction errors are also conveyed to the Actor who stores and learns an action selection policy.

The Actor uses the Critic's prediction error as a surrogate reinforcement signal with which it can improve its policy.
In this particular division of labor, the "environment" sees only the output of the Actor, (i.e., the actions), however, rewards from the environment are only of interest to the Critic.

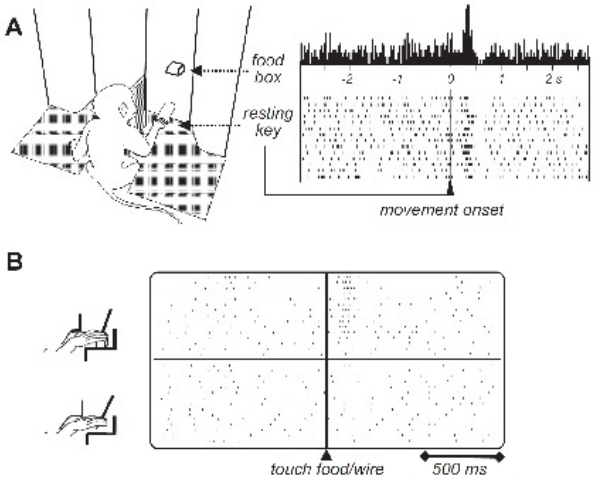
Dopamine neurons show phasic excitatory and inhibitory responses to different events. This response can be understood as a reward prediction error similar to the Rescorla-Wagner and TD prediction error.

Activation by rewarding stimuli

About 75% of dopamine neurons show phasic activation when animals touch a small piece of hidden food, or when drops of liquid are delivered to the mouth outside of any task.

Their responses distinguish rewards from non-reward objects

A number (14%) of dopamine neurons show also phasic activations when aversive stimuli are administered, such as pain pinch or electrical shock.



A) Dopamine responses to touch of food without any phasic stimuli predicting the reward. The food inside the box is invisible but is touched by the hand underneath the cover.

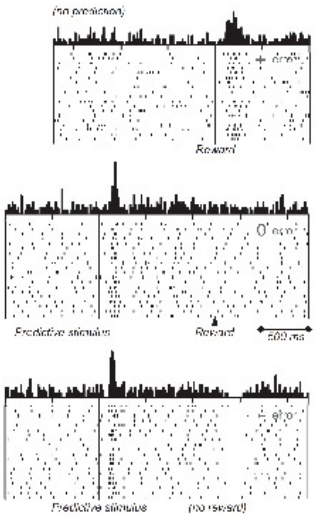
B) Differential response of dopamine neuron to touch of a wire holding a piece of apple (top), or touch of a wire holding an inedible objects (bottom).

Thus, dopamine neurons respond differentially to unpredicted objects of differing reward value, which is consistent with a prediction error signal since the prediction error will be positive when a reward is delivered but not expected.

Romo & Schultz , J. Neurophysiol, 1990

Activation by conditioned stimuli (CS)

About 55%–70% of dopamine neurons are activated by visual and auditory CS predicting reward in various Pavlovian and instrumental conditioned tasks. Only 10% of dopamine neurons are activated by CS predicting aversive events.



The transfer of the dopaminergic response from rewards to the stimuli that predict them represents the foundation of temporal difference learning in which a state or an action are associated with a scalar summary of its long-run future value,

Schultz et al., Science, 1997

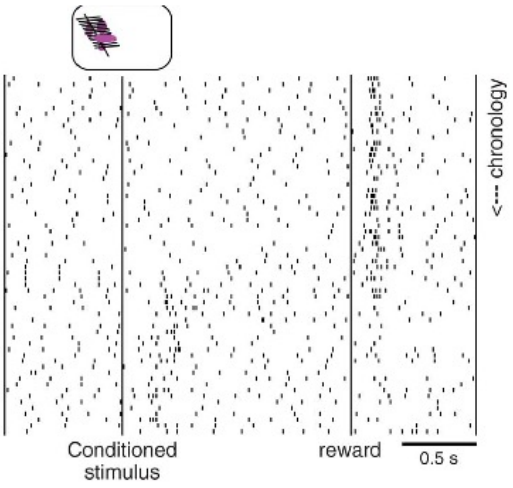
(Top panel) Before learning, a drop of appetitive fruit juice occurs in the absence of prediction—hence a positive error in the prediction of reward. The dopamine neuron is activated by this unpredicted occurrence of juice.

Middle panel) After learning, the conditioned stimulus predicts reward, and the reward occurs according to the prediction—hence no error in the prediction of reward. The dopamine neuron is activated by the reward-predicting stimulus but fails to be activated by the predicted reward (right).

(Bottom panel) After learning, the conditioned stimulus predicts a reward, but the reward fails to occur- hence a negative error in the prediction of reward. The activity of the dopamine neuron is depressed exactly at the time when the reward would have occurred.

Activation during learning

The dopamine activation undergoes systematic changes during the progress of learning. Primary rewards elicit neuronal activations during initial learning periods which decrease progressively and are transferred to the conditioned, reward-predicting stimuli with increasing learning.
During a transient learning period, both rewards and conditioned stimuli elicit an activation.



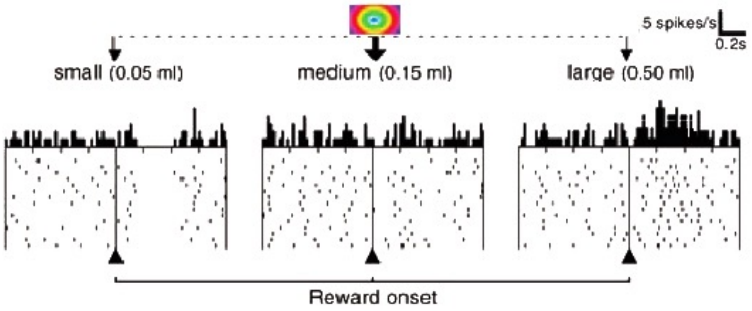
A dopamine neuron that responds initially to a juice reward acquires a response to the CS (presented on the screen) after some trials in which the CS is paired with the reward.

Chronology of trials is from top to bottom. The top trial shows the activity of the neuron while the animal saw the stimulus for the first time in its life.

Schultz, Ann. Rev. Psychol., 2006

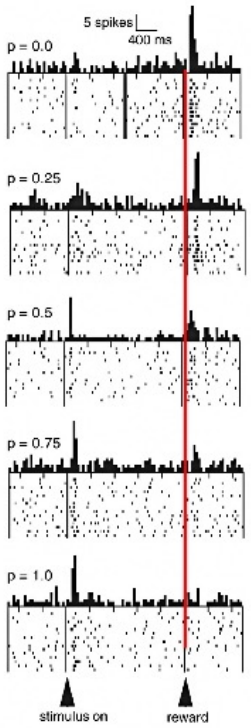
Prediction error signal is bidirectional

The neuronal response of dopamine neurons is bidirectional. A reward that is unexpectedly delivered or is better than predicted elicits an activation (positive prediction error response). An expected reward that is omitted, or is worse than predicted induces a depression (negative error response). A fully predicted reward produces no response.



After learning, a single CS was usually followed by an intermediate volume of liquid (0.15 ml) that elicited no change in the neuron's activity (center). However, on a small minority of trials, smaller (0.05 ml) or larger (0.50 ml) volumes were unpredictably substituted, and neural activity decreased (left) or increased (right), respectively.

Tobler et al., Science, 2005



Coding of reward probability by dopamine neurons

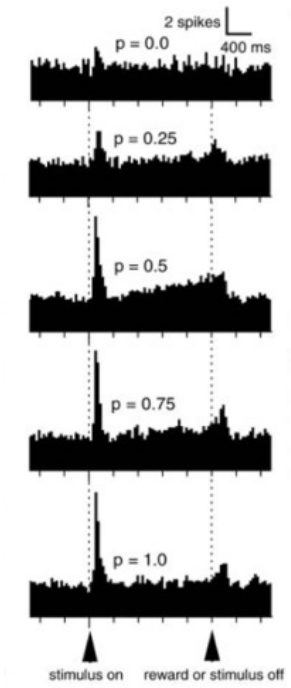
Prediction error is modulated both by predictions (expected reward) as well as (obtained) rewards.
In one study, five different visual CS predicted delivery of reward with different probabilities (p), ranging in steps of 0.25 from certain delivery (p=1), to certain nondelivery (p=0).

According to the Rescorla-Wagner and TD learning models, when the animal has learned the task, the prediction V_t for each stimulus would indicate the average reward obtained for that stimulus (e.g., 1 for the certain reward stimulus, 0.5 for the stimulus rewarded 50% of the time, and so on).

Thus, the prediction error for reward delivery would be zero for the always rewarded stimulus, and large for the never-rewarded stimulus, and something in between for the others.

Indeed, phasic dopamine responses to a reward have this property, they increase with the size of the prediction error (or, equivalently, decrease with the degree of reward probability).

Fiorillo et al., Science, 2003



Ramp-like, tonic dopamine signal encodes reward uncertainty.

Some dopamine show a sustained increase in activity that grows from the onset of the conditioned stimulus to the expected time of reward.

Furthermore, the peak of the sustained activation occurs at the time of potential reward, which corresponds to the moment of greatest uncertainty.

Tonic responses were significantly larger at p= 0.5. Common measures of uncertainty (variance, standard deviation, and entropy) are all maximal at p= 0.5.

Subjective reward uncertainty corresponds both to the necessity of identifying more accurate predictors, and to the expectation of reward information in order to resolve uncertainty

Fiorillo et al., Science, 2003

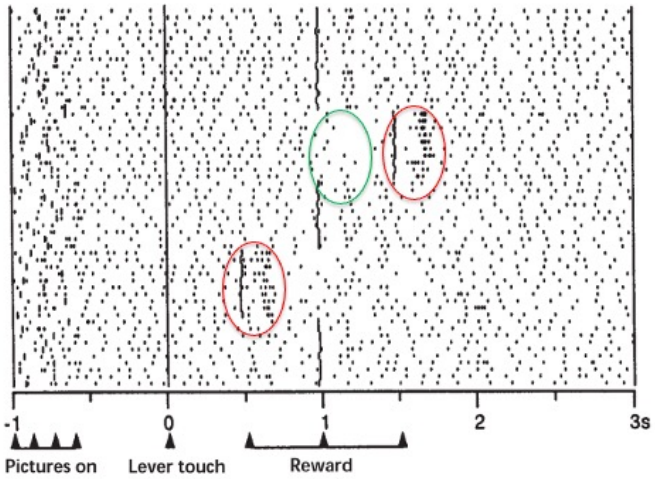
Dopamine neurons report an error in the temporal prediction of reward

Dopamine responses depend on event unexpectedness. The unexpectedness, however, is not limited to event occurrence (e.g., a reward is delivered or omitted unexpectedly), but also includes the time of reward, as rewards elicit transient activations when they are delivered earlier or later than predicted, even though it is certain that the reward will eventually occur.

Moreover, dopamine neurons are depressed exactly at the time of the usual occurrence of reward when a predicted reward is omitted.
The depression occurs even in the absence of any stimuli at the time of the omitted reward, indicating that the depression does not constitute a simple neuronal response but reflects an expectation process based on an internal clock tracking the precise time of predicted reward.

Hollerman & Schultz, Nat. Neurosci., 1998

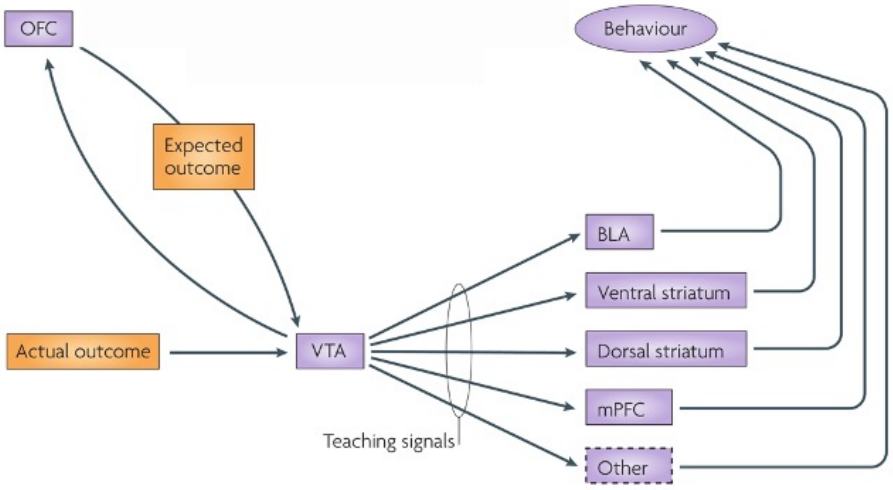
Temporal prediction error of reward



Dopamine neuron: effects of reward timing during familiar trials. Following a correct response, the reward was delivered after 1.0 s (as expected), 1.5 s (unexpected late) or 0.5 s (unexpected early). Activity of a dopamine neuron was depressed (green) when reward failed to occur at the familiar (expected) time, and increased (red) when reward unexpectedly occurred at a new time, either earlier or later.

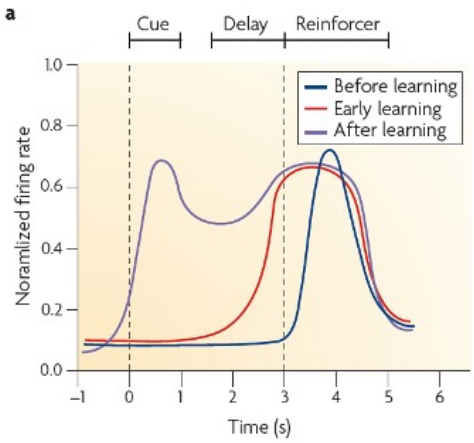
Hollerman & Schultz, Nat. Neurosci., 1998

It is commonly held that dopamine neurons broadcast a prediction error — the difference between the learned predictive value of the current state, signaled by cues or features of the environment, and the sum of the current reward and the value of the next state.



The PFC signals outcome expectancies

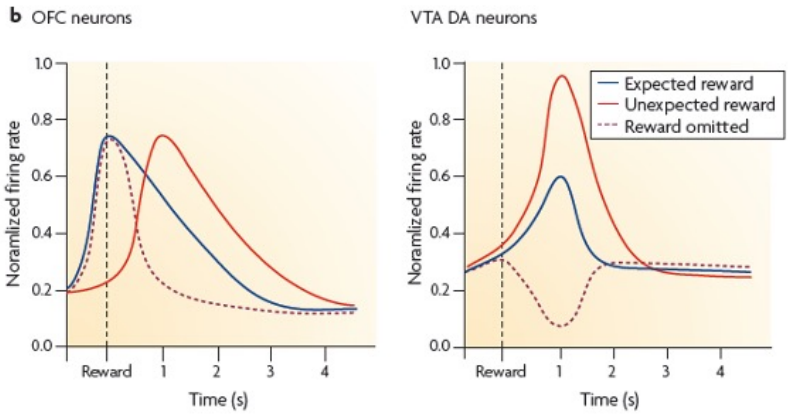
PFC signals the predicted value of specific outcomes that an agent expects given particular cues in the environment.



In associative cue-reward tasks, neurons in the PFC initially fire in response to the rewarding reinforcer (dark blue line).

After a number of trials, the neurons also start to fire in anticipation of the reinforcer (red line).

Finally, they also come to fire in response to cues that predict the reinforcer (purple line).



Moreover, PFC neurons firing is not stronger in response to an unexpected reward or weaker in response to the omission of a reward (left panel).

In this respect, PFC neurons are unlike dopamine (DA) neurons, which are also reward responsive but which fire more strongly in response to unexpected rewards and decrease firing when an expected reward is not delivered (right panel).

RPE theory of dopamine

Midbrain dopamine signals are widely thought to report RPE consistent with 'model-free' RL

- dopaminergic RPE reflects the value of observable state, that is a quantitative summary of future reward, irrespective of either the specific identity of the reward (e.g., water, food), or the sequence of future states through which it will be obtained.

- RPE theory of dopamine also posits that state values are learned through direct experience.

- dopamine signals PE for reward not for events that are surprising but not directly rewarding (or aversive);

- RPE signals should not reflect inferences based on models of the environment or task contingencies;

Dopamine encodes model-based predictions

Standard (e.g., model-free) reinforcement learning occurs through prediction errors derived from actually experienced outcomes.

However, individuals also learn about contexts, rules, and task structure, which may be called models (e.g., cognitive map) of the world.

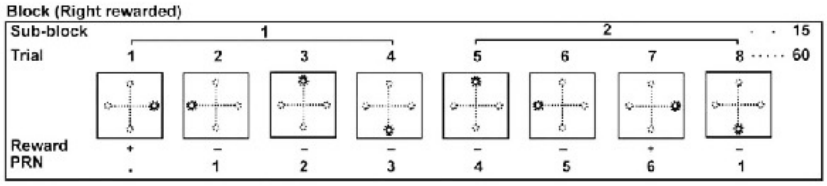
Knowledge derived from such models can deeply affect the prediction in the neuronal prediction error computation, resulting in a more appropriate teaching signal, and thus tremendously improve learning.

The acquisition and updating of some of the models likely involves cortical (for instance, in PFC), rather than dopamine signals.

In contrast, dopamine responses seem to incorporate the predictive information from models once they are established.

In one such study (Nakahara et al., 2004), sequences of non-rewarded trials lead to higher reward probability with increasing non-rewarded trials (increasing PRN, Post-Reward trial Number).

Nakahara et al., Neuron, 2004



The task was a memory-guided saccade task with four possible target positions, but reward was given to only one of these positions. A cue stimulus indicated the saccade goal.
The task included consecutive sub-blocks of 4 trials each, with one trial for each direction. Within a sub-block, one out of four directions was associated with reward (reward probability =25%), while the other three directions were not rewarded.

The task induced a specific profile of reward probability in relation to the preceding trials. The probability of reward increased with the number of previous non-rewarded trials (PNR= Post reward trial number). Reward probability was the lowest (p=0.0625), if the preceding trial was rewarded (PRN: 1). In contrast, reward probability was the highest (p=1.0), if six preceding trials were not rewarded (PRN: 7).

Thus, the animal's reward prediction (e.g., reward expectancy) should increase after each unrewarded trial. As a consequence, positive prediction error (reward delivery) should decrease, and negative prediction error (reward omission) should increase after each non-rewarded trial.
In line with this reasoning, dopamine neurons show decreasing activations to reward delivery (red line) and increasing depressions to reward omission (blue line), as the number of non-rewarded trials increases .

Nakahara et al., Neuron, 2004

Due to increasing reward prediction, later reward delivery induces increasingly weaker positive prediction errors, and reward omission elicits stronger negative prediction errors.

In contrast, model-free reinforcement learning would only consider the past unrewarded trials and hence generate progressively decreasing reward prediction and an opposite, increasing pattern of prediction errors.

The observed dopamine prediction error responses are not fully explained by the previous experience with reward (model-free learning) but incorporate predictions from the task structure (model of the world).

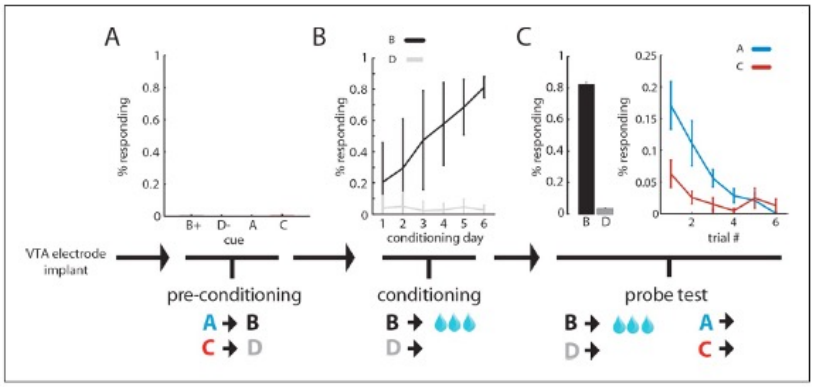
Thus, dopamine neurons process prediction errors with both model-free and model-based reinforcement learning.

Nakahara et al., Neuron, 2004

Not all dopaminergic predictions are learned through direct experience

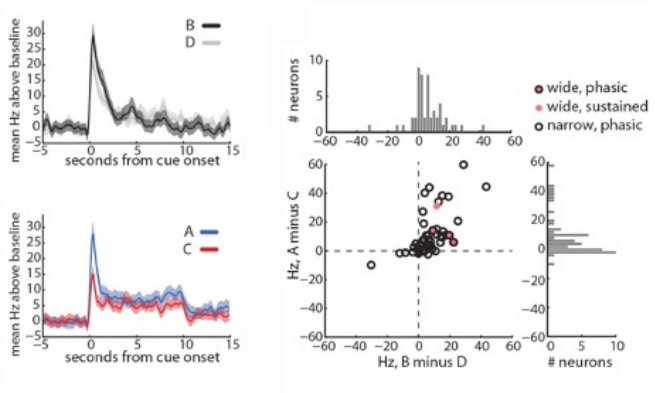
A central aspect of TDRL as model free is that values for state are learned (and cached) through direct experience with the state.

Recent work suggests, however, that phasic dopamine may reflect values that have been learned indirectly.



Sadacca et al., eLife, 2016

Dopamine neurons exhibited the largest responses at the onset of B, the reward-paired cue (significantly above responding to D), and to A, the cue that had been paired with B in the preconditioning phase (significantly above responding to control cue C).
Further, the activity elicited by these two cues was strongly correlated, suggesting that dopamine neurons code RPE errors elicited by these two types of cues in a common framework.



Sadacca et al., eLife, 2016

Dopamine RPE reflects inference over hidden-state (a belief state)

In the real world, stimuli often provide ambiguous information about states; the true underlying states are 'hidden' and must, therefore, be inferred.

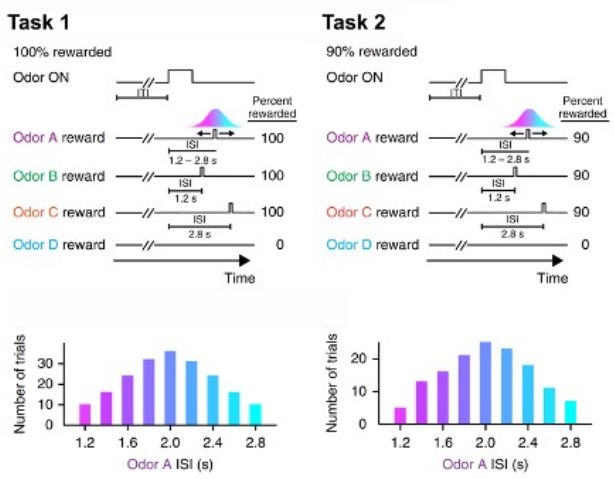
A principled way to incorporate hidden states into the TD learning framework is to replace the traditional stimulus representation with a belief state, which tracks the probability of being in each state given the trial history.

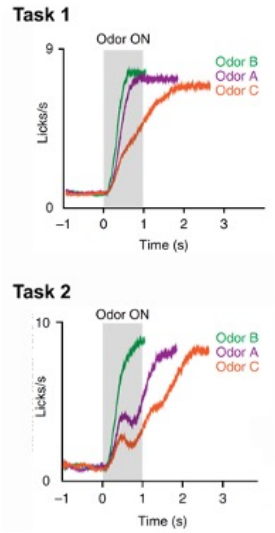
Starkweather et al tested whether dopaminergic RPEs provide evidence for a value prediction computed on a belief state.

Starkweather et al., Nat Neurosci 2017

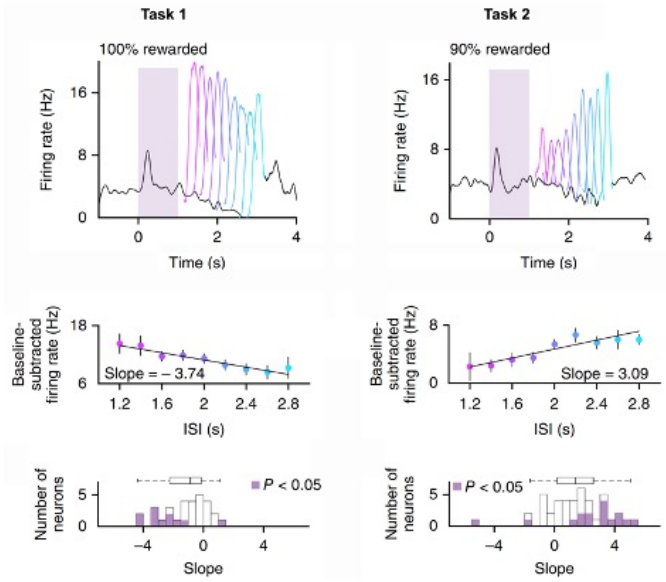
Dopamine RPE signals reflect inference over hidden-state (a belief state)

The authors elegantly demonstrate that dopamine RPE signals depends critically on the learned structure (model) of a task (i.e., whether a reward follows the cue on 100% or 90% of trials).





Starkweather et al., Nat Neurosci 2017

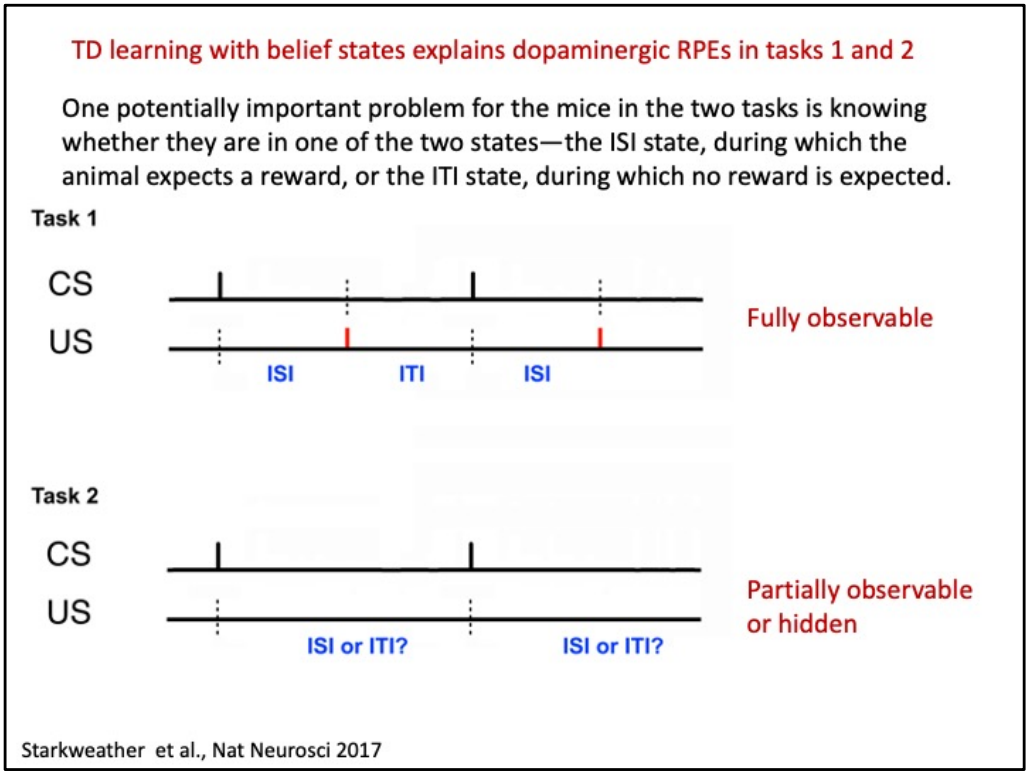


However, post-reward dopamine neurons firing showed opposite trends of temporal modulation in task 1 and 2.

In task 1, post-reward firing showed negative temporal modulation

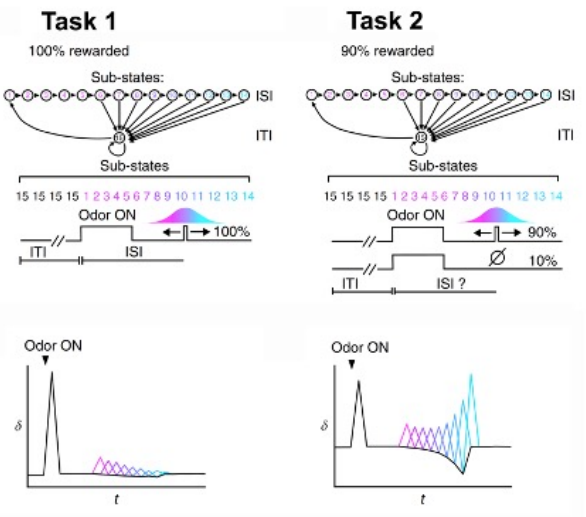
in task 2, post-reward firing showed positive temporal modulation.

Starkweather et al., Nat Neurosci 2017



In task 1, the states were fully observable, and thus the belief state was uniform throughout the variable ISI range

As soon as the cue came on, the belief state encoded a 100% probability of being in one of the ISI sub-states and a 0% probability of being in the ITI sub-state.



Starkweather et al., Nat Neurosci 2017

Task 1

The momentary probability of receiving a reward was greater at later ISIs than at earlier ISIs, resulting in RPEs that were increasingly suppressed toward the end of the variable ISI range.

Task 2

As time elapsed and the reward failed to arrive, the belief state progressively shifted in favor of the ITI state over the ISI state.
RPEs toward the end of the interval were nearly as large as the unpredicted rewards.

A long-standing idea in modern neuroscience is that the brain computes inferences about the outside world rather than by passively observing its environment.

Results showed that, depending on whether or not a reward was delivered deterministically, dopaminergic RPEs exhibited opposite patterns of temporal modulation.

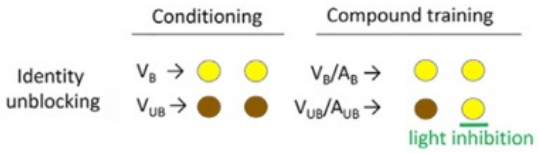
These data are well explained by a TD model incorporating hidden-state inference on the dopamine RPEs.

Dopamine as a generalized PE

Midbrain dopamine neurons are thought to drive associative learning by signaling the difference between actual and expected reward (RPE), taking into account both immediate and future rewards.

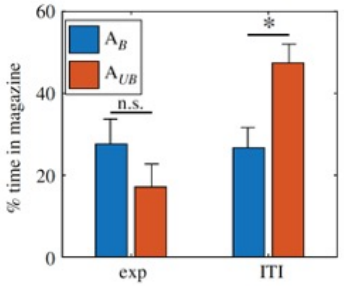
However, an increasing body of findings suggest signals carried by dopamine neurons are more heterogeneous, extending beyond rewards and rewarding cues, and including responses to aversive, motor, and cognitive variables, such as such as changes in stimulus contingencies, that should in principle be invisible to a pure 'model-free' RL system.

Dopamine is necessary for learning about changes in reward identity



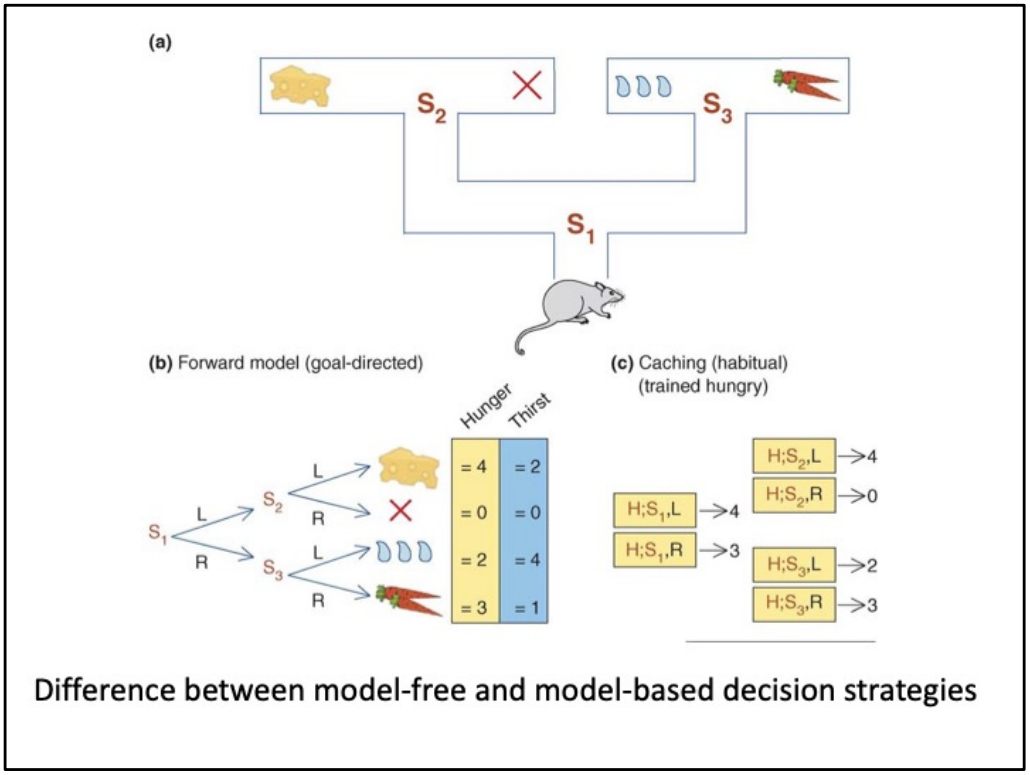
Animals first learned to associate two stimuli (V_B and V_{UB}) with different reward flavours. These stimuli were then reinforced in compound with other stimuli (A_B and A_{UB}).
Critically, the $V_{UB}A_{UB}$ trials were accompanied by a change in reward flavour, a procedure known as 'identity unblocking' that attenuates the blocking effect.

This result suggests that dopamine transients play a general role in error signaling rather than being restricted to only signaling errors in value.



Conditioned responding on the probe test. Exp: experimental group, receiving inhibition during reward outcome.
ITI: control group, receiving inhibition during the intertrial interval.

Chang et al, Curr Bio., 2017



Difference between model-free and model-based decision strategies

Two strategies to solve the sequential action selection problem.
(a) A rat navigates a maze with different outcomes at different end points. The rat starts at state S1 and must choose either left (L) or right (R). It must choose again at either S2 or S3, to turn L or R to harvest one outcome.
(b) By learning a model of the environment (consisting of a state-transition model and a reward model, essentially a state–action–outcome tree), the rat can decide whether to turn L or R at S1 by searching through the tree (simulating its next action choices) and finding the path with the highest overall utility.
Crucially, the current motivational state of the rat defines the relevant mapping between outcomes and utilities (numbers in boxes), such that when hungry (yellow), the rat will find choice L optimal at S1, but when thirsty (blue), it will prefer R. Behavior is thus goal-directed.
(c) By contrast, a model-free strategy relies on stored (cached) values in common currency for state–action pairs. These action values are estimates of the highest return the rat can expect for each action taken from each (nonterminal) state. Action selection simply involves choosing the action with the greatest cached value at the current state. Because the values are divorced from the identities of the outcomes produced by different actions, changes in the outcome–utility mapping (due to different motivation state, hunger or thirst) cannot be translated into appropriate changes in values.

However, the motivational state (hunger, H) can be stored as part of the state

representation. In this way, action selection can be modified to match a different motivational mapping (e.g. relevant to thirst,T) if the set of (state, action) values relevant to that state {(T;S1,R),(T;S2,L),. . .} has previously been learned.

Generally, when the environment of a model-free agent changes the way it reacts to the agent's actions, the agent has to acquire new experience in the changed environment during which it can update its policy and/or value function. For a model-free agent to change the action its policy specifies for a state, or to change an action value associated with a state (or both), it has to move to that state, act from it, possibly many times, and experience the consequences of its actions.
A model-based agent can accommodate changes in its environment without this kind of 'personal experience' with the states and actions affected by the change. A change in its model automatically (through planning) changes its policy.
For example, imagine that a model-based rat with a previously learned transition and reward model is placed directly in the goal box to the left of S2 to find that the reward available there now has value 1 instead of 4 (e.g., a smaller piece of cheese). The rat's reward model will change even though the action choices required to find that goal box in the maze were not involved. The planning process will bring knowledge of the new reward to bear on maze running without the need for additional experience in the maze; in this case changing the policy to right turns at both S1 and S3 to obtain a return of 3.

The Successor Representation

One possible way to reconcile the new and old findings is based on the idea that dopamine computes prediction errors over sensory features, encompassing both reward and non-reward features of a stimulus.

This sensory prediction error (SPE) can be used to estimate a predictive feature map of the environment known as the successor representation (SR),

Specifically, the SR for state X is a vector whose elements indicate the expected occupancy of future states after starting in state X.

SR separates the problem of predicting value into two components: the expected future states given the current state (i.e., the SR) and the immediate rewards available in each state.

Long-run values can be computed by multiplying the SR with the immediate reward available in each state.

$$V(s_t) = R(s_t)M(s_t, s_{t+1})$$

SR may be a middle ground between the extremes occupied by model-free and model-based algorithms

	Representation	Computation	Behavior
MF learner	*Q*: Cached value	Retrieve cached value Lowest cost	Habit, Fast
MB learner	*R*: Vector of all state rewards *T*: One-step state transitions matrix	Iteratively compute values Highest cost, resource-constrained	Fully flexible, Slow
SR learner	*R*: Vector of all state rewards *M*: Multi-step future state occupancy matrix (policy-dependent caching)	Combine cached future occupancies with rewards Intermediate costs	Semi-flexible, Fast

SR simplifies the computation of future rewards, combining the efficiency of model-free RL with some of the flexibility of model-based RL

Such a representation learning strategy is particularly well suited to environments in which the trajectories of states are fairly reliable, but rewards and goals change frequently.

Momennejad et al., Nat. Hum. Behav, 2017

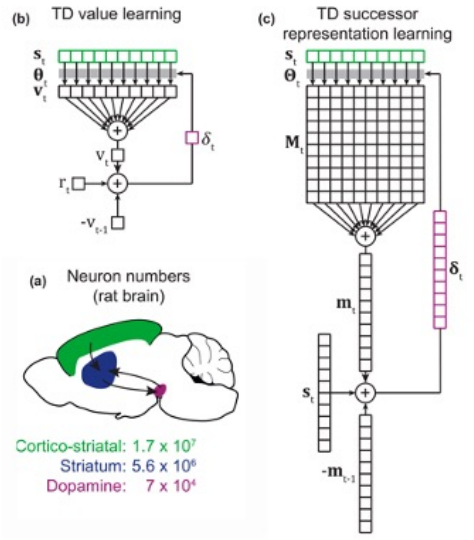
First, SR renders value computation a linear operation, yielding efficiency comparable to model-free evaluation.

Second, it retains some of the flexibility of model-based evaluation. Specifically, changes in rewards will instantly affect values because the reward function is represented separately from the SR.

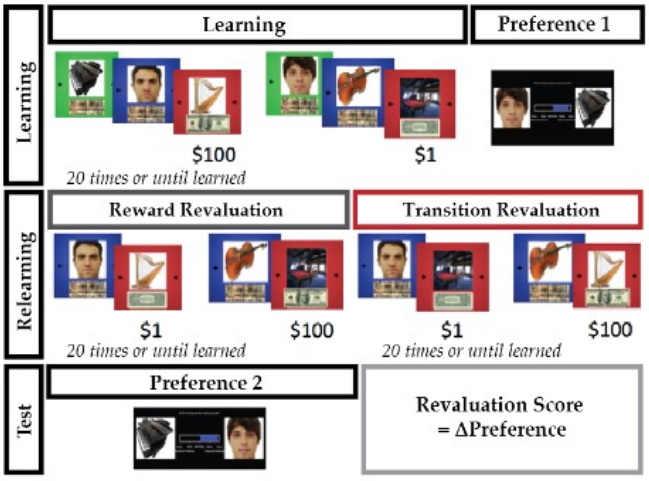
On the other hand, the SR will be relatively insensitive to changes in transition structure, because it does not explicitly represent transitions—these have been compiled into a convenient but inflexible format.

The SR caches long-term predictions about the states it expects to visit in the future. Namely, for each starting state, the SR caches how often the agent expects to visit each of its successor states in the future (learned via simple temporal difference learning).

Momennejad et al. examined whether humans employ algorithm for reinforcement learning, based on the successor representation (SR).



In phase 1 (the learning phase), participants first learned three-step trajectories leading to reward.



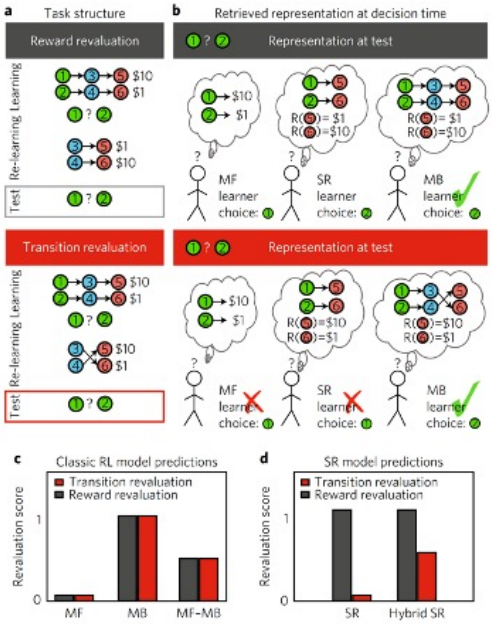
At the end of the learning phase, they were asked to indicate which starting state they believed led to greater future reward by reporting their relative preference using a continuous scale. Learning was assessed by the participant's preference for the starting state associated with the more rewarding trajectory

Momennejad et al., Nat. Hum. Behav, 2017

Participants were exposed to one stimulus at a time and were asked to indicate their preference for the middle state after every five stimuli.
The learning phase ended if the participant indicated preference for the highest paying trajectory three times, or after 20 stimulus presentations.

In phase 2 (the re-learning phase), trajectories were initiated at the middle state of the trajectory, and the structure of the task was altered in one of two ways (within participants): in the reward revaluation condition, the rewards associated with the terminal states were swapped, whereas in the transition revaluation condition, the transitions between the step 2 and step 3 states were swapped.

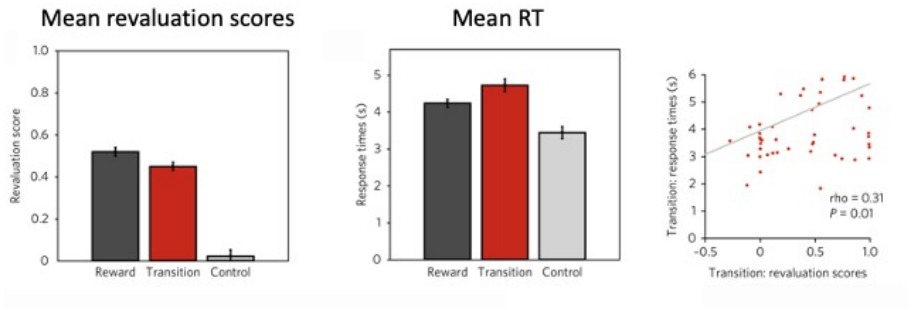
Momennejad et al., Nat. Hum. Behav, 2017



A purely SR learner retrieves the SR without further computation and readily combines it with the reward vector.

An MB learner computes full transitions (a costly computation), then combines them with the reward vector R to produce a decision

Momennejad et al., Nat. Hum. Behav, 2017

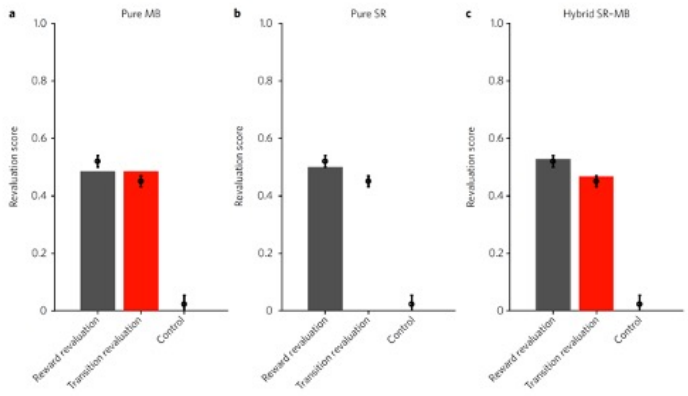


Accuracy was significantly higher in the reward revaluation than in the transition revaluation condition; RTs were faster in the reward revaluation than in the transition revaluation condition.
There was a significant correlation between the accuracy of transition revaluation responses and response times: more accurate transition revaluation took longer, suggesting that successful transition revaluation might have involved more computation at the decision time.

Momennejad et al., Nat. Hum. Behav, 2017

Human performance was measured as the change in preference ratings for the starting states. Revaluation scores for each game denote the change in a given participant's relative preference rating after versus before the relearning phase.

Together with significantly faster RTs compared with reward revaluation, the positive correlation between the accuracy of transition revaluation responses and response times
lends further evidence to the possibility that, compared with reward revaluation, transition revaluation required more cycles of computation at the decision time, relying less on cached representations.



Model fits to the phase 3 test data from the learning task.

Model performance (solid bars) against human data (error bars) using a pure MB learner (a), a pure SR learner (b) and a hybrid SR–MB learner (c). Human behaviour is best explained by the hybrid account.

Momennejad et al., Nat. Hum. Behav, 2017

Human performance was measured as the change in preference ratings for the starting states. Revaluation scores for each game denote the change in a given participant's relative preference rating after versus before the relearning phase.

Together with significantly faster RTs compared with reward revaluation, the positive correlation between the accuracy of transition revaluation responses and response times
lends further evidence to the possibility that, compared with reward revaluation, transition revaluation required more cycles of computation at the decision time, relying less on cached representations.

While MB performs equally well on all revaluation tests and MF solves none, the SR can use its cached representations to readily solve reward revaluation, but not transition revaluation.

SR is a predictive model of the environment, which allows the mental simulation of distal future events rapidly.

It differs from the one-step model representations used in standard MB learning, mainly because it aggregates these predictions over many future time steps.

Distributional Reinforcement Learning in the brain

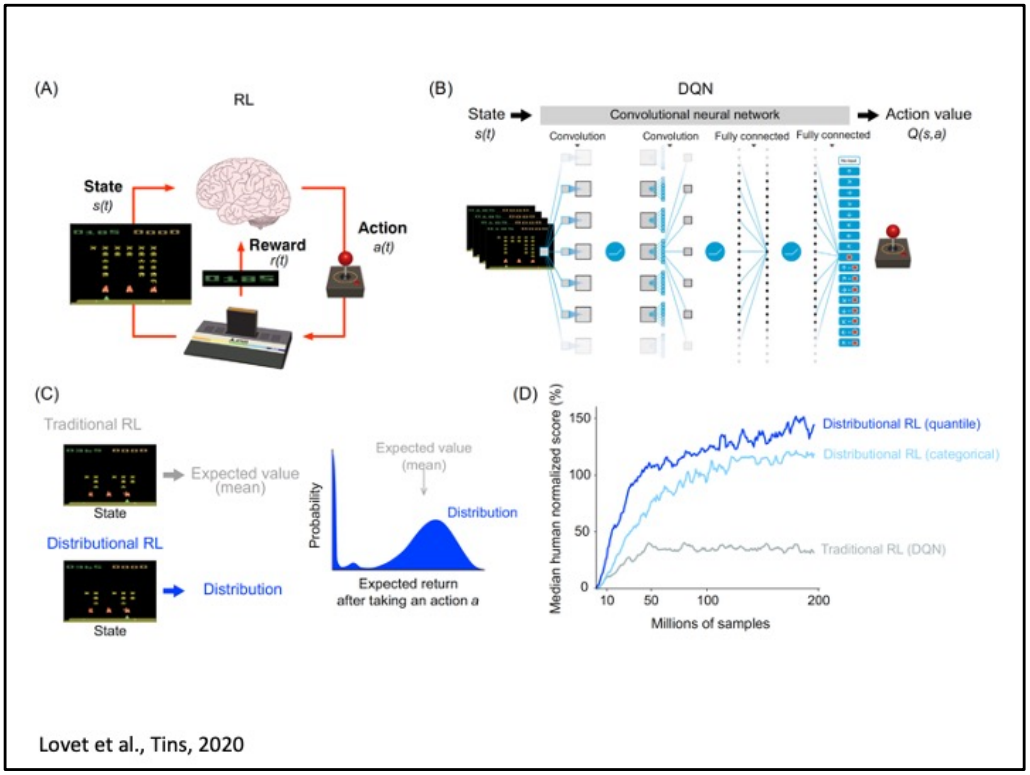
Recent advances in RL highlights the importance of learning the full distribution of returns over expected (mean) return.

Reward distribution contains more information about the consequence of the agent's decision.

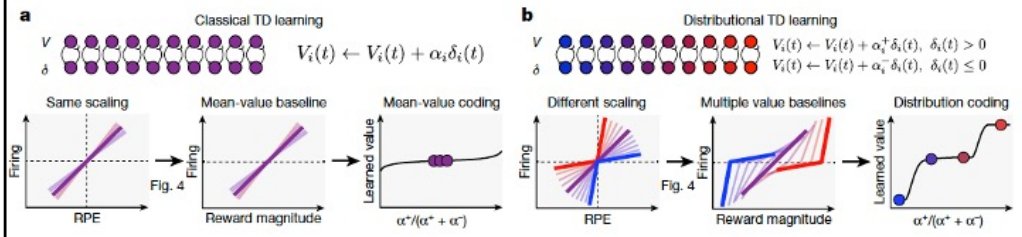
Compared with classical RL procedures, distributional RL can increase performance in deep Q-learning systems by a factor of two.

This prompted the question of whether RL in the brain, the dopamine neurons particularly, might leverage the benefits of distributional coding.

Indeed, the brain utilizes distributional codes in numerous other domains, including sensory and cognitive domain.



Lovet et al., Tins, 2020



Distributional RL posits multiple value predictors and RPE channels, each of which carries a different prediction error.

These predictors vary in their degree of optimism about upcoming reward. More optimistic predictions anticipate obtaining greater future rewards; less optimistic predictions anticipate more meager outcomes.

As a surprising consequence, a single reward outcome can simultaneously elicit positive RPEs (within relatively pessimistic channels) and negative RPEs (within more optimistic ones).

Dabney et al., Nature, 2020

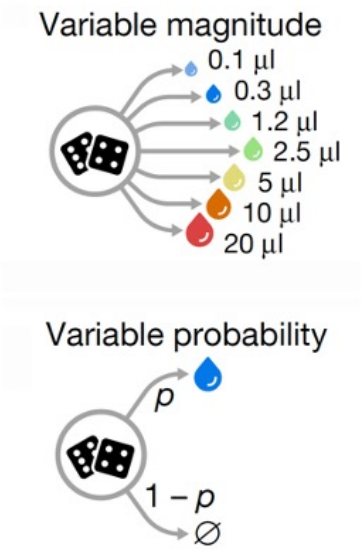
DA neurons have different reversal points

For a given DA neuron, let us call reversal point the amount of reward r0 for which, if a reward $r < r0$ is received, the neuron expresses a negative error, and if a reward $r > r0$ is received, it expresses a positive error.

Under the traditional TD learning model, individual neurons should show approximately identical reversal points and should weigh positive and negative errors equally.

However, experimental evidence suggests otherwise

Dabney et al., Nature, 2020



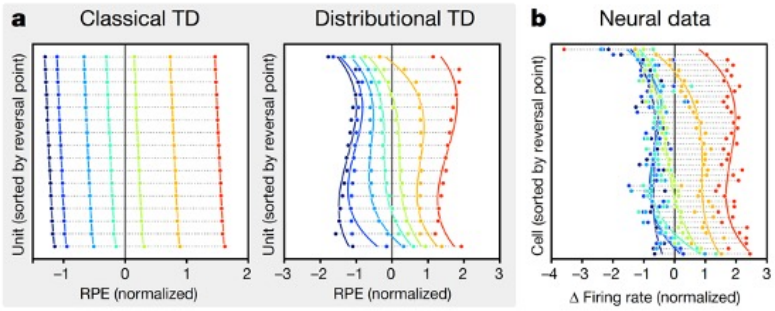
Mice were trained on a 'variable-probability' task, and different mice on a 'variable-magnitude' task.

In the variable-magnitude task, one of the following reward magnitudes was delivered, at random: 0.1, 0.3, 1.2, 2.5, 5, 10 or 20 µl. In half of these trials, this reward was preceded by 1,500 ms by an odour cue (which indicated that a reward was forthcoming but did not In the other half, reward was unsignalled.

In the variable-probability task, in each trial animals experienced one of 4 odour cues for 1s, signalling 90%, 50%, 10% or 0% chance of reward (3.75 µl water).

Dabney et al., Nature, 2020

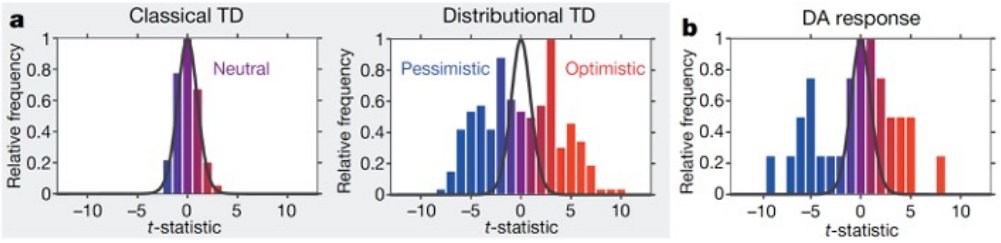
Different dopamine neurons reverse from positive to negative responses at different reward magnitudes (variable magnitude task)



Each horizontal bar is one simulated (a) or recorded (b) neuron. Each dot colour corresponds to a particular reward magnitude. The x axis is the cell's response when reward is delivered. Cells are sorted by reversal point.

In classical TD, all cells carried approximately the same RPE signal.
Conversely, in distributional TD, cells had reliably different reversal points.
Some responded positively to only the very largest reward (top optimistic neuron) and other responded positively to almost all rewards (bottom pessimistic neuron).

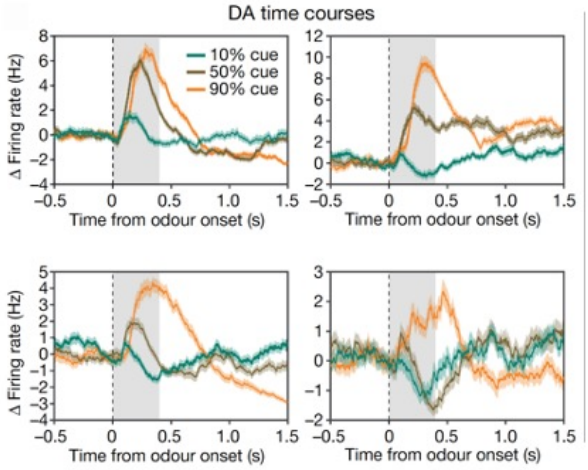
Dabney et al., Nature, 2020



Data from variable-probability task.
Histogram (a, simulated cells; b, and recorded DA neurons) of t-statistics which compare each cell's 50% cue response against the mean 50% cue response across cells.

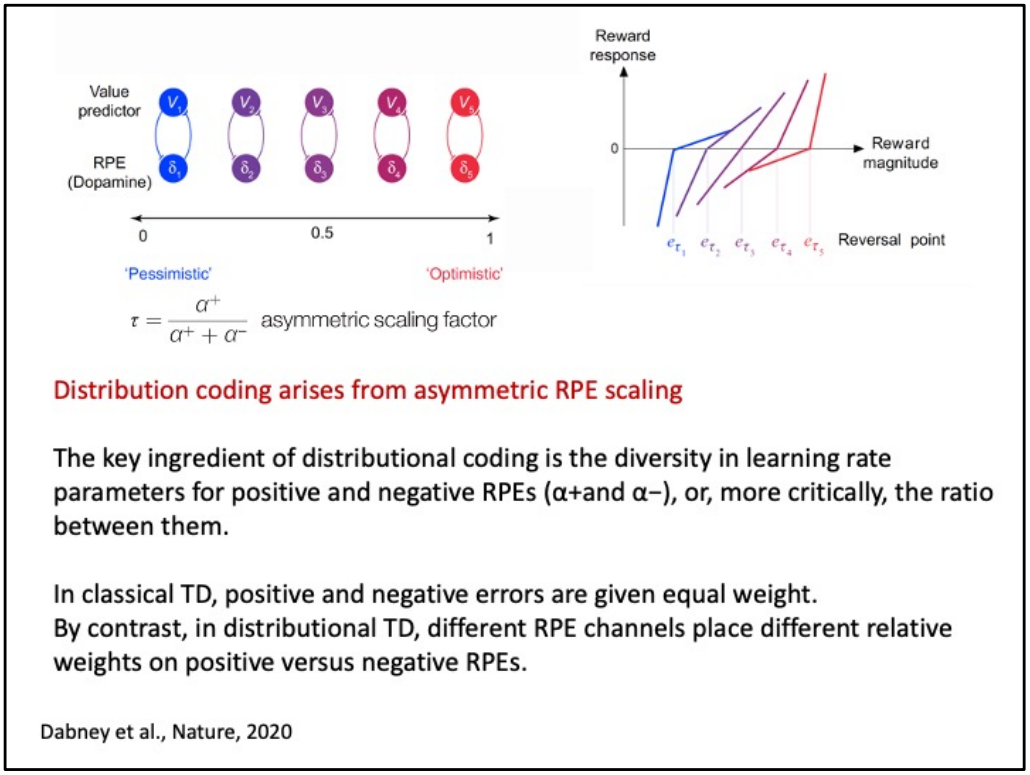
In Classical TD, no difference is expected when comparing the 50% cue response against the midpoint of 10% and 90% responses.
Distributional RL predicts, instead, that dopamine neurons should vary in their responses to the 50% cue

Dabney et al., Nature, 2020



Responses of four example dopamine neurons recorded in a single animal
Some neurons (top) respond optimistically, emitting a RPE nearly as large as to the 90% cue, in their responses to the 50% cue.
Others respond pessimistically (bottom), emitting a RPE closer to the 10% cue response, in their responses to the 50% cue.

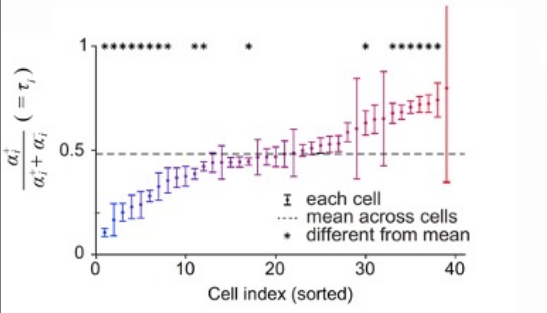
Dabney et al., Nature, 2020

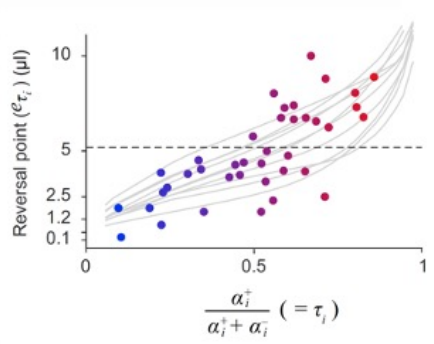


In classical TD, positive and negative errors are given equal weight, as a result, positive and negative errors are in equilibrium
Therefore, classical TD learns to predict the average over future rewards.

By contrast, in distributional TD, different RPE channels place different relative weights on positive versus negative RPEs.
In channels that overweight positive RPEs, reaching equilibrium requires these positive errors to become less frequent, so the learning dynamics converge on a more optimistic reward prediction. Conversely, in channels overweighting negative RPEs, a more pessimistic

prediction is needed to attain equilibrium.





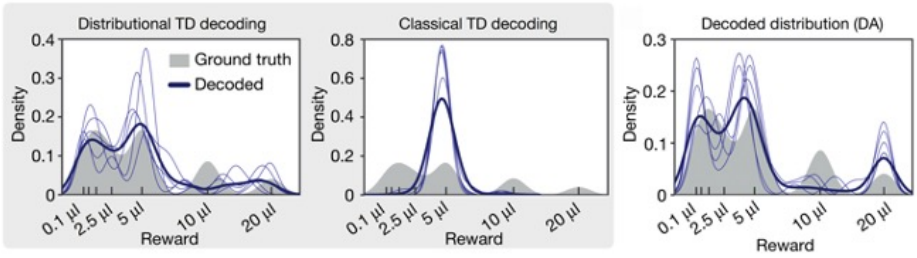
Across all animals, the mean value of the ratio α+/(α+ + α−) was 0.48. However, many cells had a value significantly above or below this mean.

Using data from the variable-magnitude task, a strong correlation was found between RPE asymmetry and reversal point: Dopamine neurons that scale positive RPEs more steeply relative to negative RPEs have reversal points at relatively high reward magnitudes. Conversely, dopamine neurons that scale positive RPEs less steeply have relatively low reversal points.

Dabney et al., Nature, 2020

For each cell, two slopes: α+ for responses in the positive domain (that is, above the reversal point), and α− for the negative domain were separately estimated.

Decoding reward distributions from neural responses



Distributional TD trained on the variable-magnitude task (left, center)

Dabney et al., Nature, 2020