OPINION

Believing in dopamine

Samuel J. Gershman and Naoshige Uchida

Abstract | Midbrain dopamine signals are widely thought to report reward prediction errors that drive learning in the basal ganglia. However, dopamine has also been implicated in various probabilistic computations, such as encoding uncertainty and controlling exploration. Here, we show how these different facets of dopamine signalling can be brought together under a common reinforcement learning framework. The key idea is that multiple sources of uncertainty impinge on reinforcement learning computations: uncertainty about the state of the environment, the parameters of the value function and the optimal action policy. Each of these sources plays a distinct role in the prefrontal cortex–basal ganglia circuit for reinforcement learning and is ultimately reflected in dopamine activity. The view that dopamine plays a central role in the encoding and updating of beliefs brings the classical prediction error theory into alignment with more recent theories of Bayesian reinforcement learning.

The neuromodulator dopamine lives a double life. On the one hand, it is thought to convey the discrepancy between observed and expected reward, known as the reward prediction error (RPE), which serves as a learning signal for updating reward expectations in the striatum[1,2]. On the other hand, it appears to participate in various probabilistic computations, including the encoding of uncertainty and the control of uncertainty-guided exploration. The purpose of this Opinion article is to bring together these different roles into a common reinforcement learning framework.

The key ingredients of reinforcement learning theories are state, value and policy. In reinforcement learning, values are computed on the basis of the state of the world that the animal currently occupies. A state is collectively defined by the animal's location, the time from certain events, what objects are present and so on. The value of a state is defined as the discounted sum of all future rewards starting from the state. A policy is the function that determines which actions are selected in each state. Our starting point is the recognition that animals face several forms of uncertainty encompassing all of these ingredients — state, value and policy.

First, animals commonly do not have full information about which state they are currently occupying. Rather, they receive sensory data that provide ambiguous information about the current state[3,4]. For example, an animal might sense an odour plume to infer the hidden location of a food source. Because many different locations could be compatible with the odour plume to various degrees, the normatively correct strategy is to compute the posterior probability distribution of the food location conditional on the odour information. This computation is stipulated by Bayes' rule: $P(\text{location}|\text{odour})$ is proportional to the product of the likelihood $P(\text{odour}|\text{location})$ and the prior $P(\text{location})$.

Second, animals must learn a mapping from states to predictions about future rewards (the value function). For example, a foraging animal must learn how much cumulative food it can expect to collect by foraging in a particular patch. When the state space is large, an approximation of the value function is typically specified by a set of parameters (for example, the value of a patch is approximated by a weighted sum of its features, such as its size and resource density). Because these parameters are unknown, the animal has uncertainty about them, which is gradually resolved through the experience of rewards in different states. Whereas standard models of learning, such as the temporal difference (TD) model, update point estimates of the parameters (that is, a single set of parameter values), other models encode uncertainty about the parameters in the form of a probability distribution over the parameters[5,6].

Third, animals must compute a mapping from states to action probabilities (the policy). This mapping is typically mediated by learned values, such that actions that take the animal to rewarding states tend to be selected. However, since the optimal policy is unknown, animals must balance the need to exploit actions with known rewards against the need to explore actions that might potentially have better rewards (the exploration–exploitation dilemma). Intuitively, uncertainty should motivate exploration: an animal should gather information about actions to reduce uncertainty about their values. Two forms of uncertainty-guided exploration have been the subject of recent studies[7–10]. One approach is to add an 'uncertainty bonus' to the learned values, such that actions are biased to explore unfamiliar actions (directed exploration). Another approach (random exploration) is to increase the stochasticity of the policy in proportion to uncertainty.

We argue that these three forms of uncertainty (associated with states, values and policies) exert distinct effects on midbrain dopamine activity by impinging on different stages of the information processing architecture for reinforcement learning (FIG. 1). As we elaborate herein, these effects can be formalized in terms of Bayesian reinforcement learning principles. The Bayesian framework significantly enriches the traditional RPE interpretation of dopamine signalling, allowing it to accommodate a broader range of phenomena, and leading to new predictions that have recently been tested experimentally. The framework also delineates the computational functions of the medial prefrontal cortex (mPFC) and orbitofrontal cortex, and how they interact with the dopamine system. Finally, we discuss how this framework embraces a role for dopamine in the encoding of policy uncertainty and the control of exploration.

State uncertainty

Consider the problem faced by a foraging animal in the African savannah (FIG. 1): whether to forage or not to forage in a particular patch of grass depends on whether the animal believes that a lion is hiding in the grass. Because its sensory data provide ambiguous information about the hidden state (that is, the presence or absence of a lion), the normatively correct representation of uncertainty is the posterior probability distribution over the hidden state conditional on the sensory data, which can be computed using Bayes' rule. There is abundant evidence that animals represent posterior probability distributions[11]. From a reinforcement learning perspective, the question is how the animal should use the posterior probability distribution to predict future rewards and ultimately select a reward-maximizing action.

An elegant solution to this problem is provided by the concept of a belief state[12,13]. As mentioned earlier, the environmental state is a sufficient statistic for reward prediction: if the animal knows what state it is in, it can optimally predict future rewards without needing to store its state history in memory[14]. This 'memoryless' property of the reinforcement learning problem is what makes possible efficient algorithms, such as dynamic programming (value iteration) and TD learning (BOX 1). When the state is hidden, the problem is no longer memoryless, because the optimal estimate of the state depends on the entire history of past observations. However, the agent need not store this entire history in episodic memory; the posterior probability distribution encodes all the available information for predicting future reward. This distribution can thus be regarded as a 'state' in the sense that it is a sufficient statistic for reward prediction. We will henceforth refer to the posterior probability distribution as the belief state.

The belief state plays the same role as other state representations in the standard reinforcement learning machinery. Specifically, a value function maps the belief state to an estimate of cumulative future reward, which may be conditioned on action to support downstream decision computations. Importantly, this mapping may be learned via dopamine RPEs, and hence these signals should reflect the underlying belief state. Later, we unpack each step of this machinery as it applies to belief states. First, we describe how belief states are computed in the mPFC. Second, we describe how the striatum encodes belief states using a set of basis functions, which are then mapped to values. The encoding step allows the striatum to selectively retain information about the belief state that is useful for predicting reward. Finally, we describe how midbrain dopamine neurons compute RPEs from the striatal value estimates.

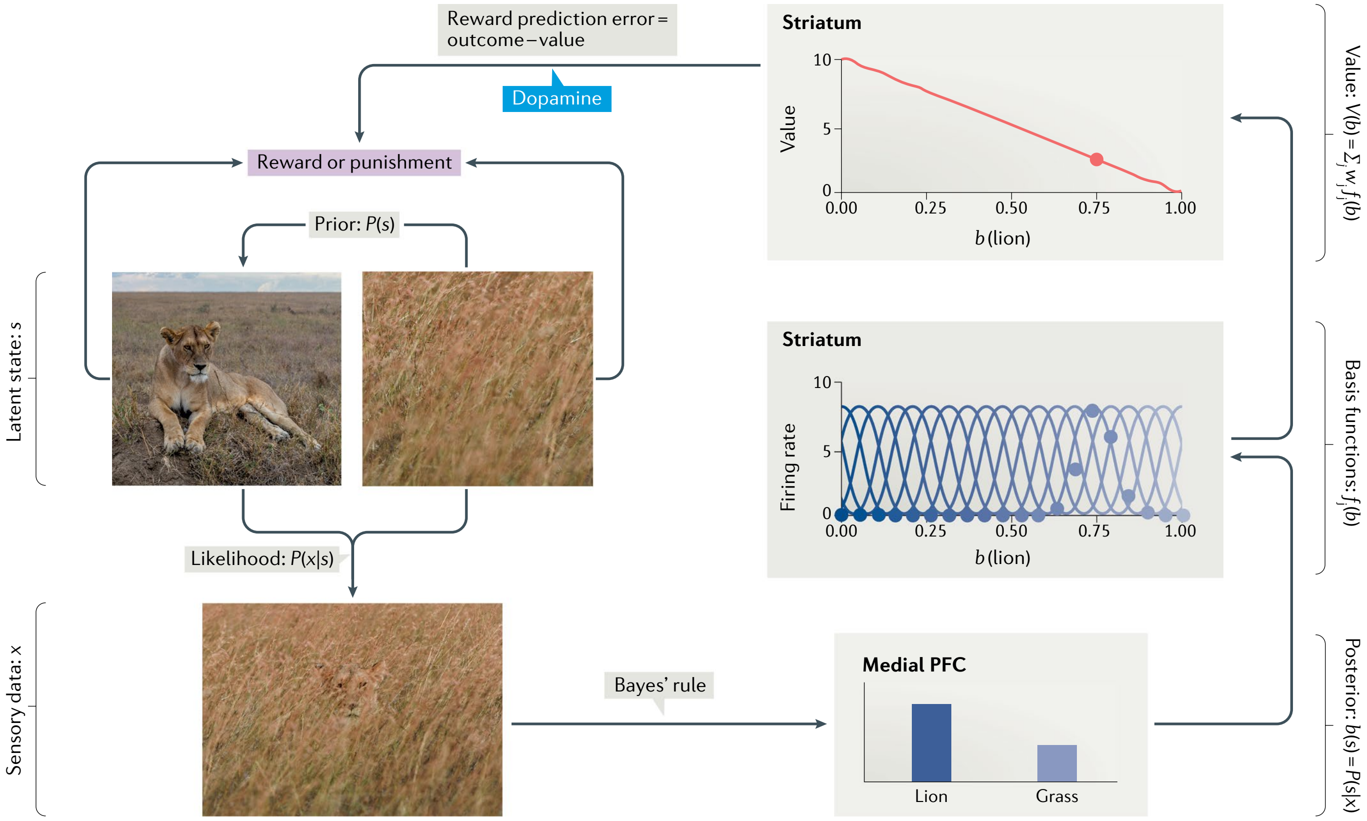

Fig. 1 | **The neural architecture for reinforcement learning under state uncertainty.** Bayesian inference combines noisy and ambiguous sensory data with a prior over latent states to compute the posterior probability distribution, or belief state, hypothesized to be encoded in the medial prefrontal cortex (PFC). For example, a retinal image of grass may contain ambiguous information about the presence of a lion (this hidden state). The likelihood encodes the degree to which a hypothetical hidden state predicts the sensory data (what is the evidence for a lion hiding in the grass?), and the prior distribution encodes the probability of hidden states before the sensory data have been collected (how frequently do lions hide in this grass?). The belief state is mapped into a distributed state representation (basis functions) in the striatum, which is in turn mapped onto a value function (an estimate of expected future reward). Dopamine drives the updating of the value function parameters by reporting a reward prediction error (the difference between observed and expected reward, or value).

Box 1 | Temporal difference learning with belief states

Most reinforcement learning algorithms, including temporal difference learning[112], assume that the environment can be described by a Markov decision process, consisting of a state transition function, $T(s'|s,a)$, specifying the probability of transitioning from state s to state s' after taking action a, and a reward function, $R(s)$, specifying the expected reward in state s. This generative model of the environment obeys the Markov property: state transitions and rewards are independent of the agent's history conditional on the current state.

When the state is hidden, the environment is typically modelled as a partially observable Markov decision process[112,113], which additionally includes an observation function, $O(x,s)$, which specifies the probability of observing sensory data x in state s. Under partial observability, the environment is no longer Markovian in the sensory data: future observations are not independent of the agent's history conditional on the current observation. However, the environment is Markovian in the posterior probability distribution over states, $b(s)$, which can be computed from the sensory data using Bayes' rule:

$$b(s) = P(s|x) \propto O(x, s)P(s)$$

where $P(s)$ is the prior over states.

The temporal difference algorithm can be applied directly to the belief state representation using a standard linear function approximation:

$$V(b) = \sum_j w_j f_j(b)$$

$$\Delta w_j = \alpha\ \delta f_j(b)$$

where $f_j(b)$ is a basis function (indexed by j) over belief states, w_j is the coefficient associated with basis function j, α is a learning rate and δ is the reward prediction error:

$$\delta = r + \gamma V(b') - V(b)$$

where r is the reward received in state s and γ is a discount factor that exponentially down-weights future rewards. Although we have used a linear function approximation for clarity, nonlinear approximations are also possible.

Belief state representation. When the state space is discrete (or suitably discretized), the belief state corresponds to a vector of probabilities (the probability of being in each state), which could be directly encoded by the firing rate of individual neurons or populations of neurons[15]. One limitation of such a 'labelled-line' code is that the number of neurons required scales exponentially with the dimensionality of the state space. This limitation can be addressed by optimizing a parametric approximation of the exact posterior probability distribution[16,17] or by approximately sampling from the posterior probability distribution to construct a non-parametric approximation[18–21].

Several lines of evidence point to the mPFC as a candidate locus for belief state computation. Changes in mPFC activity track the updating of the posterior probability distribution[22–26], and damage to the mPFC is associated with aberrant belief formation, such as confabulation[27]. Changes in mPFC activity ('network resets') are also associated with the onset of behavioural variability[28], consistent with the idea that the variance of the posterior probability distribution (representing the animal's uncertainty) controls the randomness of the action policy, as discussed further later[8,28].

Another line of evidence comes from reversal learning experiments, in which two or more reward contingencies alternate. Animals become progressively faster at adapting to these reversals ('rapid reacquisition'), in some cases requiring only a single trial to dramatically change their behaviour[29–32], a phenomenon that is inconsistent with models of learning in which reward predictions are relearned after each reversal. As several authors have noted[29,33], reversal learning is better modelled as a problem of hidden state inference: each reward contingency corresponds to a hidden state, and the animal normatively should combine ambiguous reward information with its prior over hidden states via Bayes' rule. In addition, it must simultaneously estimate the parameters governing each state. As the animal becomes increasingly confident in its estimates of these parameters, it will be better able to identify reversals and hence switch more rapidly. Lesions of macaque mPFC appear to leave rapid reacquisition intact, but they increase the rate of reversal, consistent with either a reduced evidence threshold or an increased estimate of the reversal probability[34]. A functional MRI study of reversal learning in humans[35] found that mPFC activity is sensitive to hidden state inference.

Although we have focused on the mPFC, belief updating is probably distributed across many different brain regions, depending on the task, input modality and other factors. It is currently unclear whether reinforcement learning circuits receive preferential input from one or more belief-encoding regions.

Value function approximation. From a reinforcement learning perspective, the goal of a belief state computation is ultimately to support reward prediction and control. An exact representation of beliefs may be computationally wasteful if rewards can be predicted accurately from a lower-fidelity representation. Moreover, even if computational resources were not limited, an ideal agent would still need to restrict the space of value functions that map belief states to rewards, because a more complex class of value function is more likely to overfit the data. One standard way to accomplish this restriction is to approximate values as linear functions of a set of basis functions that are computed from the state representation. Although a more complex nonlinear value function approximation is possible[36], most models of the basal ganglia assume a linear function approximation architecture.

Following an earlier proposal[12], we hypothesize that the striatum encodes the set of basis functions. These basis functions can be thought of neurally as cells that are tuned to particular regions of the belief space (the 'belief points'). Presently, the existence of basis functions defined over belief states is still speculative (indeed, the nature of striatal basis functions more generally is shrouded in mystery), but there are some suggestive pieces of evidence, primarily from tasks involving timing uncertainty.

Many tasks require animals to estimate elapsed time, and it is well known that timing uncertainty increases with interval duration, a property known as scalar timing[37]. Pacemaker–accumulator models[38] explain this phenomenon mechanistically: an accumulator noisily counts pulses emitted from a pacemaker, and these counts are compared with a reference retrieved from memory, corrupted by multiplicative noise. From a Bayesian perspective[39,40], the hidden state corresponds to elapsed time, the prior distribution corresponds to the reference memory (the set of likely interval durations) and the likelihood corresponds to the accumulator process (the evidence accrued for a particular interval). One implication of the scalar property is that the posterior probability distribution over elapsed time (the belief state) will be broader for longer intervals. If this belief state is represented by cortical inputs to the striatum using a labelled-line code, then each cortical neuron is tuned to a particular hidden state (elapsed time) and its firing rate is proportional to the posterior probability of that hidden state. The width of the population is broader when timing uncertainty is greater (that is, for longer intervals)[41].

If we assume that striatal basis function neurons receive input from a subpopulation of similarly tuned cortical belief state neurons, then the temporal profile of striatal activation will be more spread out for longer intervals, owing to the broader width of the cortical population code (BOX 2). We could equivalently conceptualize the striatal neurons as tuned to elapsed time, with receptive fields that broaden for longer intervals. This is precisely the idea put forth by the microstimulus model[42–44], which has successfully explained a range of data on dopamine physiology and classical conditioning.

Approximately Gaussian-shaped temporal receptive fields have been reported in rodent striatum[45–48] and primate striatum[49], whereas other studies in rodents have reported monotonic tuning (that is, ramping) in this region[50]. Consistent with a causal role for striatal time cells in downstream computations, the temporal specificity of both behaviour[51] and dopamine activity[52] depends on the integrity of the striatum.

In sum, the data from interval timing experiments are broadly consistent with a set of striatal basis functions defined over temporal belief states, but little is known about whether this generalizes to other kinds of state spaces, such as spatially[53,54] or visually[55,56] defined states.

Belief-dependent reward prediction errors in Pavlovian conditioning. If value functions are computed from belief states, then RPEs should be modulated by belief. This hypothesis was originally put forth by theorists seeking to account for experimental deviations from the predictions of the standard TD model, which assumes a fully observable state[12,13,57]. For example, in one study[58], monkeys were shown two boxes, one of which always contained food and one of which never contained food. When the door to the food-containing box opened, the firing rate of dopamine neurons increased, as expected from the standard TD model[2]. However, the firing rate also increased when the door to the other (no-food) box opened, contradicting the standard TD model, according to which only reward-predicting cues will elicit a positive RPE. A similar finding was reported in another study that explicitly manipulated the reward context: dopamine responses generalize to unrewarded stimuli when they occur in the same context as rewarded stimuli[59]. More recently it was found that this reward generalization can apply even to aversive stimuli[60]. One possible explanation for these findings is that the monkey is initially uncertain about which box is going to open, and therefore its value estimate defined on the belief state would reflect a mixture of the two box-specific values, producing a positive RPE[2,57]. This explanation also accounts for another feature of the data: a suppression of dopamine activity immediately after the burst in response to the no-food box opening. According to the belief state model, the value goes from positive to zero once the state uncertainty is resolved, and hence the TD prediction error is negative.

More detailed predictions have been derived for Pavlovian conditioning tasks in which the presentation of a cue is followed by a reward after some delay (the interstimulus interval (ISI)). The delay between the reward and the onset of the next cue is the intertrial interval (ITI). Daw et al.[13] modelled this task as consisting of separate ISI and ITI states, parametrized by a dwell-time distribution (determining how long each state is occupied), a transition distribution (determining which states are visited after the dwell time has elapsed) and a reward distribution (determining how much reward is delivered in each state). Formally, this corresponds to a semi-Markov process (BOX 2). If the reward is delivered stochastically, then the state becomes hidden, because the animal does not know whether the absence of reward signals an omission trial or a transition to the next ITI.

In this belief state model, a reward delivered earlier than expected will result in a positive prediction error, just as in the standard fully observable TD model. However, the models make different predictions about what will happen at the expected time of reward. The standard model predicts a negative RPE, because the expected reward has been omitted. By contrast, the belief state model predicts that the animal will infer a transition to the next ITI, and hence its reward expectation will go to zero. This prediction is consistent with empirical observations from studies with monkeys: no suppression of dopamine activity is observed at the expected time of reward[61].

Recent studies with mice have built on these findings, pursuing a more detailed empirical test of the belief state model's

Box 2 | A unifying view of state representation

The 'standard' temporal difference model applied to dopamine signalling uses a complete serial compound (CSC) representation of time, which represents each stimulus as a collection of binary features, each of which is 'on' after a specific delay following stimulus onset. Formally, $f_j(t) = 1$ exactly j time steps after the onset of the stimulus, where t indexes time. In essence, the CSC chops poststimulus time into a collection of discrete bins and attaches a separate coefficient w_j to each bin. Neurally, the CSC can be implemented using a set of stimulus-tuned and temporally tuned neurons. Although simple and widely used, the CSC has been criticized for making incorrect predictions[13,42,114,115].

Two alternative representations have played an important role in recent theorizing. One alternative, based on a semi-Markov model, replaces the discrete time bins with a continuous representation of dwell time[13]. For example, in a Pavlovian conditioning task, an animal enters the interstimulus interval state when the conditioned stimulus appears, and this state may be occupied for a random dwell time. Although this state representation seems quite distinct from the CSC, one can use the CSC to construct a discrete-time Markov approximation of the semi-Markov dynamics[62], in which the time bins correspond to 'substates'; the key difference from the standard temporal difference model is that the transition probabilities between substates are chosen to match the dwell time distribution, rather than proceeding ballistically after stimulus onset.

Another alternative replaces the uniform-width time bins with 'microstimulus' basis functions, the width of which increases and amplitude of which decreases as a function of time[42,43,62]. The discrete-time Markov approximation of the semi-Markov model offers one way of deriving these basis functions. If the uncertainty about the substate grows as a function of time, then the belief state will become increasingly spread out across multiple substates, exhibiting the same qualitative properties as microstimuli. A key difference is that the temporal profile of belief states depends on the task structure. The observation that time cells in the striatum (putative microstimulus-like basis functions) rescale under different fixed interval schedules suggests that the basis functions are adaptive[45].

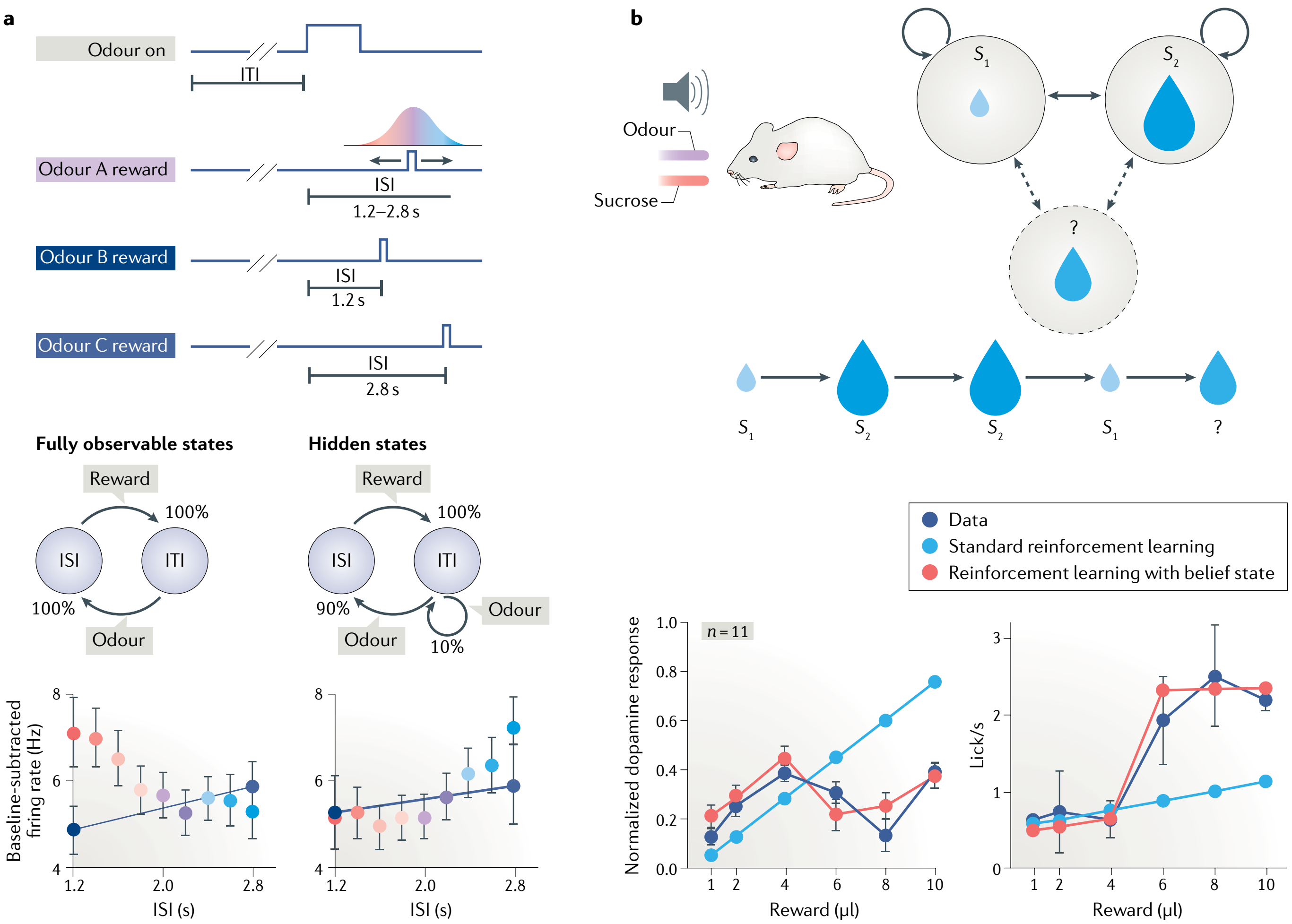

Fig. 2 | **Experimental evidence for reflections of state uncertainty in dopamine signals. a** | Mice observed an odour followed by a water reward. Odour A was associated with a variable odour–reward interval, whereas odours B and C were associated with fixed intervals. The middle plots show the structure of the task as a probabilistic graphical model. The bottom plots show the baseline-subtracted firing rates of optogenetically identified dopamine neurons in the ventral tegmental area. Firing rates decreased with the odour–reward interval when the reward was deterministic, but they increased with the interval when the reward was probabilistic, consistent with the predictions of the belief state reinforcement learning model. **b** | Mice observed an odour followed by a water reward the magnitude of which varied across blocks. The middle plot shows the normalized calcium response from dopamine neurons in the ventral tegmental area measured using fibre photometry. The bottom plot shows anticipatory licking and the predicted values for both a standard reinforcement learning model (no belief state) and a belief state reinforcement learning model. Mice were trained using blocks of either small or big reward trials first. In rare trials (probe trials), mice received intermediate-size reward. The *x*-axis indicates the magnitudes of reward in the probe trials. ISI, interstimulus interval between odour and reward; ITI, intertrial interval. Part **a** is adapted from REF.[62], Springer Nature Limited. Part **b** is adapted from REF.[66], CC-BY-4.0

predictions (FIG. 2a). When rewards are delivered deterministically, there is a monotonic decrease in the response of dopamine neurons to reward delivery as a function of ISI[62,63]. Because there is no state uncertainty under deterministic reward delivery (the animal always knows that it is in the ISI until the reward is delivered), the belief state and standard TD models both correctly predict that reward expectation will grow as a function of ISI and hence that the RPE will decrease. When rewards are delivered stochastically (10% of rewards are omitted), the pattern changes radically: dopamine neuron activity increases as a function of ISI. The belief state model, but not the standard TD model, predicts this finding as a consequence of the fact that as the ISI grows, the animal will become increasingly confident that a state transition has occurred, causing the reward expectation to decrease and the RPE to increase. Several additional analyses also ruled out an alternative account based on subjective hazard functions[63,64].

Consistent with the hypothesized role for the mPFC in state inference, the effect of state uncertainty on dopamine neuron responses is disrupted by inactivation of this brain region[62,65]. Specifically, the monotonic increase in the dopamine response as a function of ISI in the 90% reward condition flattens with mPFC inactivation. Importantly, there is no effect of mPFC inactivation on the monotonic decrease in response in the 100% reward condition. Furthermore, the sensitivity of other dopamine neuron responses to interval timing (for example, a ‘dip’ during reward omission) remains intact. These results suggest that the mPFC is specifically involved in belief-dependent RPEs when there is state uncertainty.

Another recent study tested the belief state model’s predictions using a novel variant of reversal learning[66] (FIG. 2b). Mice first alternated between two conditions distinguished only by the reward magnitude.

On small reward blocks, animals received an odour cue and then shortly afterwards a small water reward. Large reward blocks were identical (including the same odour cue), except that the water reward magnitude was 10 times larger. After training, mice exhibited anticipatory licking (a proxy for value) that scaled with reward magnitude. The small and large conditions continued in a test phase, but occasionally the mice would receive a block in which rewards were of an intermediate magnitude. On these intermediate blocks, the belief state model asserts that RPEs will be a non-monotonic function of reward magnitude. Intuitively, small intermediate rewards provide evidence that the mouse is in the small reward state, and because the mouse is receiving a reward that is greater than expected in the small reward state, the RPE should be positive. As the intermediate reward increases, the RPE will increase correspondingly. However, when the intermediate reward reaches the midpoint between the small and large rewards, the mouse will switch to believing that it is more likely to be in the large reward state, at which point it is receiving less than expected, producing a large negative RPE. The size of this RPE will diminish as the intermediate reward continues to increase.

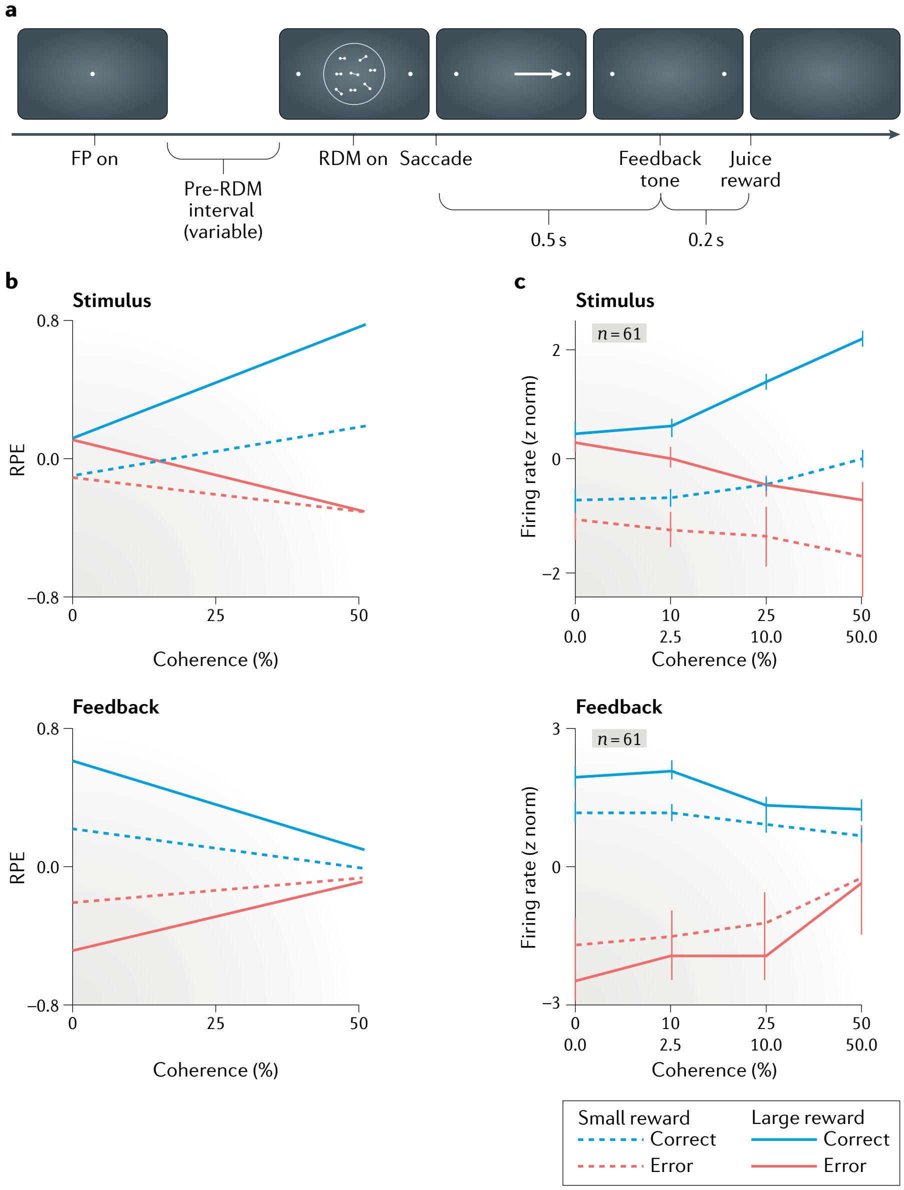

Fig. 3 | **Experimental evidence for uncertainty-dependent dopamine signals in a perceptual decision making task. a** | Monkeys observed random dot motion (RDM) and then made a saccadic decision about the overall direction[68,69]. If the decision was correct, the monkeys were rewarded with juice. The proportion of coherently moving dots was manipulated across trials. **b** | The belief state model predicted that at stimulus onset, reward prediction error (RPE) response would increase as a function of coherence on correct trials but decrease as a function of coherence on error trials. This pattern would be inverted at feedback onset. For both stimulus and feedback onset, the pattern would be amplified for larger rewards. **c** | Recordings of midbrain dopamine neurons under the same condition as in part **b** confirmed the theoretical predictions. FP, fixation point. Part **a** adapted with permission from REF.[67], SFN (https://www.jneurosci.org/content/30/32/10692). Part **b** adapted with permission from REF.[68], Elsevier.

Dopamine neuron responses conformed to this predicted 'zigzag' pattern. The same pattern was also reflected in the animals' anticipatory licking behaviour: changes in lick rate from one trial to the next tracked the RPE, as we would expect from the learning equations, and these changes were non-monotonic functions of reward magnitude. Moreover, when the belief state model was fit to the dopamine response for each individual mouse, the same model could accurately predict mouse-specific variations in anticipatory licking (despite not being fit to the behaviour). The standard TD model, by contrast, can predict only a monotonic pattern of dopamine responses (no zigzag) when using a single hidden state for all blocks, and could not as accurately predict variations in anticipatory licking.

Belief-dependent prediction errors in perceptual decisions. Another line of evidence for the belief dependence of dopamine signalling comes from perceptual decision making tasks. In the most extensively analysed study, midbrain dopamine neurons were recorded while monkeys performed a random dot motion discrimination task[66,67]. On each trial, the monkeys saw a set of moving dots, with some proportion of the dots (the coherence) moving either left or right. The monkeys reported perceived direction by a saccade to one of two targets. A key finding from this study was that the size of the stimulus-evoked dopamine neuron response correlated with coherence. This is predicted by the belief state model, because the RPE at the time of stimulus onset should be equal to the value associated with the stimulus, and higher coherence predicts higher future reward[12]. Concomitantly, the RPE at the time of reward delivery should be greater for low coherence, because the expected reward is lower on those trials, consistent with the empirical data. A recent reanalysis of these data further verified critical predictions of the belief state model[68]. Conventional TD models reflect stimulus–reward associations as mentioned earlier and predict that the stimulus-evoked dopamine neuron response should not be modulated by the animal's choice. By contrast, a belief state TD model uses the inferred stimulus, which drives the animal's choice and, at the same time, modulates the animal's confidence about receiving a reward. Supporting the belief state TD model, the reanalysis showed that the dopamine response depends jointly on performance (correct versus incorrect) and coherence[68] (FIG. 3).

Related results have been found with a vibrotactile detection task[69]. Monkeys judged whether a weak vibrotactile stimulus occurred or did not occur during an observation period. On hit trials (in which monkeys correctly detected the stimulus), dopamine neuron responses to the stimulus increased monotonically with stimulus amplitude, which can be thought of as analogous to coherence in terms of its effect on the belief state. Furthermore, dopamine neuron responses were higher on false alarm trials (in which monkeys incorrectly reported the stimulus when it did not occur) than on correct rejection trials (in which monkeys correctly reported that the stimulus did not occur). This finding indicates that dopamine neuron responses reflect subjective beliefs about the stimulus rather than the objective stimulus, consistent with the belief state model.

Box 3 | Kalman temporal difference learning

Here, we present a simplified version of the Kalman temporal difference model[45,71]. The posterior over function approximation weights is Gaussian with mean $\hat{w}$ and covariance matrix Σ. Similarly to the standard temporal difference model (BOX 1), the Kalman temporal difference model updates the mean using a reward prediction error signal:

$$\Delta w_j = \alpha_t \bar{\delta}$$

where $\alpha = \Sigma \cdot h$ is a vector of learning rates corresponding to the projection of 'temporal difference features' $h = x - \gamma x'$ onto the posterior covariance matrix Σ. The reward prediction error $\bar{\delta} = \delta / \lambda$ is normalized by the marginal variance $\lambda = h^{\mathrm{T}} \cdot \alpha + \sigma^2$, where σ^2 is the variance of the reward distribution. This hypothesized normalization is consistent with data indicating that uncertainty rescales reward prediction errors[116]. A greater unpredictability of rewards will increase λ and thus decrease the learning rate, whereas greater subjective uncertainty about the weights (encoded by α) will increase the learning rate, consistent with studies showing that high volatility of cue-reward associations leads to faster learning[117].

The posterior covariance is updated according to

$$\Delta\Sigma = \Sigma + Q - (\alpha \cdot \alpha^{\mathrm{T}}) / \lambda$$

where Q is the covariance of the weight dynamics. This model can be equivalently implemented using a recurrent neural network that transforms the temporal difference features using linear attractor dynamics. These dynamics asymptotically decorrelate the feature space. The recurrent weights can be learned using a form of anti-Hebbian learning.

Value uncertainty

In the models described so far, a parameterized value function was defined over a belief state and the parameters were estimated using RPEs. These models represent uncertainty about states but not about the value function parameters. In principle, an agent can also have uncertainty about these parameters; technically, this would correspond to treating the parameters as part of the hidden state[70]. One analytically tractable and neurally plausible special case is Kalman TD learning (BOX 3), which closely resembles the classical learning algorithm applied to phasic dopamine[2] but has the additional advantage of dynamically tracking uncertainty. This allows the TD model to connect with the substantial literature indicating that animals use uncertainty to guide learning[3,5] and to explain some puzzling properties of dopamine activity[71].

Behavioural evidence for value uncertainty. One of the classic pieces of evidence for error-driven learning comes from 'blocking' experiments. In such experiments, an animal first learns to associate a stimulus (A) with a reward (A+). Then, the stimulus is paired with another stimulus (B) while the animal continues to be rewarded (AB+). In the final phase, the animal is tested on the second stimulus without reward (B−). Despite the fact that B is consistently paired with reward, the test phase typically reveals a weak or absent conditioned response; evidently, the association between A and the reward 'blocks' the learning of an association between B and the reward[71,72]. These findings indicate that the correlation between presentation of a stimulus and receipt of a reward is not a sufficient condition for learning.

The Rescorla–Wagner model[73] offered what came to be the most influential explanation of blocking: learning is driven by prediction errors, and because A reliably predicts reward, there is no residual error to drive learning about B on the compound training trials (that is, on AB+ trials). This explanation of blocking is inherited by TD learning, and is supported by the observation that the dopamine response to AB+ is suppressed in the blocking procedure[73,74]. Moreover, blocking can be counteracted by optogenetic stimulation of midbrain dopamine neurons during compound training[75].

Despite the elegance of this account, it fails to explain why reversing the order of A+ and AB+ phases also — albeit under more restrictive conditions — produces a blocking effect on B (so-called backward blocking, to contrast it with the forward blocking effect described in the preceding paragraph)[75–77]. During AB+ training, there should be a positive prediction error to drive learning about B, since A has not yet been reliably paired with reward on its own. Somehow, training with A+ causes the association between B and the reward to be modified, a process that is not allowed under the Rescorla–Wagner model or the standard TD model; in these models, errors can drive learning of only present stimuli. An answer to this problem is provided by a Bayesian treatment of the TD model, known as Kalman TD learning[5,78] (BOX 3), which retains the successful elements of the standard TD model but also allows learning about absent stimuli. The key idea is that the reward expectation for the compound stimulus (AB) cannot exceed the sum of the expectations for A and B individually. This means that there must be a negative covariance between the stimulus-specific expectations: when the expectation for A increases during A+ trials, the expectation for B must decrease, thus producing backward blocking[79]. Mechanistically, the negative covariance corresponds to inverting the sign of the learning rate for absent stimuli.

This idea has broad applicability beyond backward blocking. Many learning phenomena involve 'retrospective revaluation' conditions, in which training appears to alter the reward expectations for absent stimuli[80]. For example, in the forward blocking paradigm, extinguishing A following compound training increases the response to B in the later, test phase[78]. Another application of the Kalman

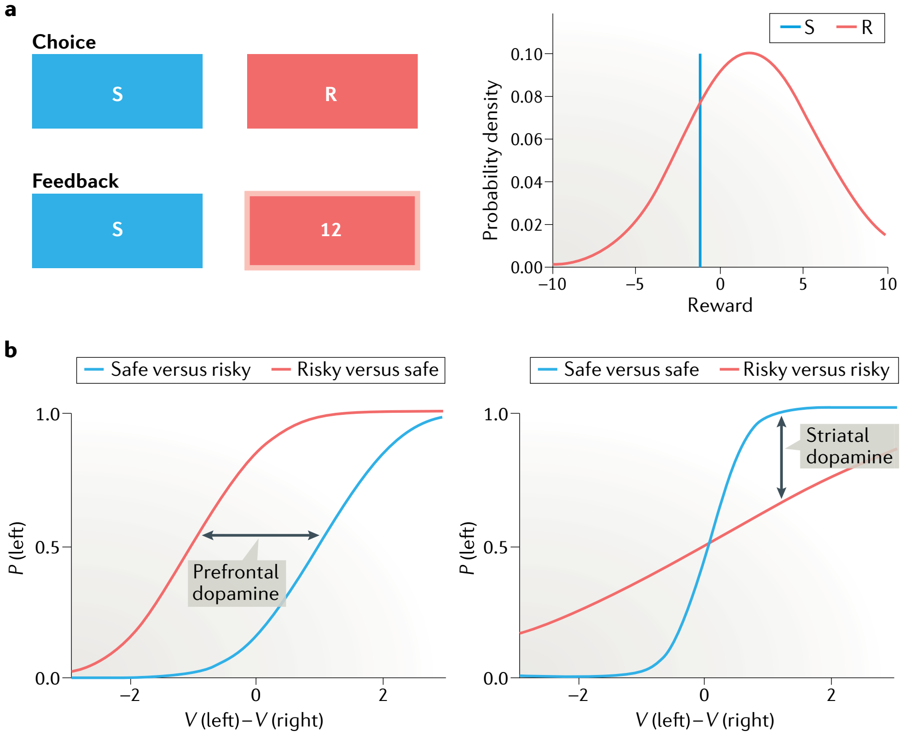

Fig. 4 | **Two forms of uncertainty have distinct effects on exploratory choice and are governed by distinct dopamine afferents. a** | A two-armed bandit task in which each arm is either 'safe' (deterministic) or 'risky' (stochastic). After choosing an arm, human participants observe a scalar reward signal superimposed on the arm chosen. The probability distribution on the right shows the payoff probabilities for the safe (S) and risky (R) arms. **b** | How different trial types affect the probability of choosing the left option, plotted as a function of the estimated value difference between the options. The left plot illustrates the manipulation of relative uncertainty: when the left option is safe and the right option is risky, the choice probability function is shifted to the right, reflecting a change in choice bias (indifference point) caused by an uncertainty bonus for the risky option. This corresponds to a form of directed exploration, putatively controlled by prefrontal dopamine levels. Evidence suggests that the magnitude of the uncertainty bonus is controlled by prefrontal dopamine levels. The right plot illustrates the manipulation of total uncertainty: when both options are safe, the choice probability function becomes steeper relative to when both options are risky, reflecting a reduction in choice stochasticity. This corresponds to a form of random exploration, putatively controlled by striatal dopamine levels. Part **a** is adapted with permission from REF.[94], APA.

TD model is to single-cue learning: when a neutral stimulus is pre-exposed (presented repeatedly without reward), subsequent conditioning of that stimulus takes longer, a phenomenon known as latent inhibition[81]. In the absence of reward, the standard TD model predicts no learning during the pre-exposure phase. The Kalman TD model, by contrast, incrementally reduces its value uncertainty during pre-exposure, becoming more confident that the stimulus predicts no reward. More training during the subsequent conditioning phase is required to overcome this belief[5]. The model also explains why interposing a delay between pre-exposure and conditioning attenuates the latent inhibition effect[82]. Under the assumption that values change gradually over time, the delay will inflate uncertainty about value, counteracting the effect of pre-exposure.

Reflections of value uncertainty in dopamine. If the Bayesian interpretation of retrospective revaluation is correct, then we should expect to see this credit assignment process reflected in dopamine signals. A case in point comes from a study of sensory preconditioning in rats[83]. In the preconditioning phase, stimulus A was paired serially with stimulus B. Note that because no reward was delivered in this phase, the standard TD model does not predict any learning. The Kalman TD model, by contrast, will learn a positive covariance between A and B, because the offset of A is associated with the onset of B[71]. In the second phase, B was paired with reward and, finally, in the third phase, the response to A was probed[83]. Behaviourally, the rats showed a conditioned response to A, even though A was never paired with the reward. This finding is consistent with the Kalman TD model, which predicts that the positive covariance will drive generalization of reward expectation from B to A. The Kalman TD model also correctly predicts that dopamine neurons will reflect this generalization, responding to A more than to a control stimulus that underwent preconditioning but lacked a second-order association with reward.

Dopamine measurements, by slow microdialysis methods in aversive conditioning, have provided some support for the Bayesian interpretation of latent inhibition described above. The stimulus-evoked dopamine level in nucleus accumbens during the conditioning phase was reduced following pre-exposure in rats[84], consistent with the assumption that RPEs will propagate more slowly to the cue onset for the pre-exposed cue[71]. More experiments are needed to confirm the generality of these findings.

Stimulus transformation in the orbitofrontal cortex underlying value uncertainty. The Kalman TD model can be implemented in a neural circuit that uses recurrent inhibition to project the 'raw' state representation onto the posterior covariance matrix (BOX 3). This transformed representation can then be linearly mapped to reward predictions, and the posterior probability distribution over the parameters of this mapping can be updated using dopamine RPEs. A candidate locus for this transformation process is the orbitofrontal cortex, which may have the appropriate network architecture[85,86] and has been implicated in state representation more broadly[87]. Consistent with this hypothesis, neurons in the orbitofrontal cortex in rats come to reflect the associative structure of sensory preconditioning[88] and lesions of the orbitofrontal cortex impair the sensory preconditioning effect[87,89].

Policy uncertainty

Ultimately, the brain's reinforcement learning system is designed not just to estimate values but to identify the optimal policy. In this section, we discuss several approaches to this problem and the putative role of dopamine.

Uncertainty-guided exploration. Several studies have found evidence for an 'uncertainty bonus' in human exploratory choice[8,9,90]. Specifically, options associated with greater uncertainty receive a bonus that is added onto the option's estimated payoff. When this bonus is larger, the policy will tend to be more exploratory. Uncertainty bonuses are one way of implementing an uncertainty-directed exploration strategy. There is also evidence that humans increase the variability of choice in proportion to their uncertainty[8,91]. A classic example of such 'random exploration' is the payoff variability effect: choices are more variable when rewards are more variable[92,93].

Recent studies have shown that these strategies can be simultaneously identified in human choice behaviour[8] and can be manipulated orthogonally[94]. Directed exploration is sensitive to the relative uncertainty between options. This is easiest to conceptualize when there are two options that have the same average payoff but one has a more variable payoff. In this case, the relative uncertainty will be non-zero, and this will induce a preference for the more variable option. Thus, relative uncertainty can be manipulated by comparing conditions in which one option is risky (variable payoffs) and the other option is safe (deterministic payoffs), or vice versa. Random exploration is sensitive to total uncertainty across the options. In the two-option case, this uncertainty is greatest when both options are risky.

When the value difference between the options is varied, we can plot choice probability as a function of the value difference, and this provides a geometric interpretation of directed and random exploration (FIG. 4). Relative uncertainty changes the intercept (indifference point) of the choice probability function, whereas total uncertainty changes the slope of the choice probability function. By fitting psychometric functions to choice behaviour using probit regression, we can extract directed and random exploration effects from the estimated coefficients.[8,94]

Using this method, a recent study showed that single-nucleotide polymorphisms in two dopamine genes were differentially involved in directed and random exploration. Variation in *COMT*, which primarily controls prefrontal dopamine levels, was selectively associated with directed exploration, confirming the results of an earlier study[95]. Variation in *DARPP32* (also known as *PPP1R1B*), which primarily controls striatal dopamine levels, was selectively associated with random exploration, consistent with prior biophysical modelling[95,96] (although this modelling work did not directly simulate the effects of *DARPP32* variations).

Dopamine as precision under active inference. The uncertainty-guided exploration strategies described are simple and effective heuristics for approximating the computationally intractable optimal solution to the exploration–exploitation dilemma. A different line of theoretical work has attempted to derive principled heuristics from the free-energy principle, which states that brain function is organized to reduce expected surprise. Applied to action selection, the imperative to reduce surprise leads to active inference: actions should be selected that fulfil the predictions of a generative model[7,17]. At first glance, this seems to be in direct opposition to the principle that actions should be taken to gain information about the world (for example, sensory predictions could be trivially fulfilled by sitting in a dark room), but critically the free-energy principle assumes that the generative model also optimizes a probability distribution over motivational states, such as hunger, as well as more abstract hierarchies of goals[97]. Hunger, according to this analysis, is 'surprising', in the sense that it violates a prior belief that hunger states should be unlikely. When surprise is minimized over longer time horizons, active inference will select actions that not only reduce immediate hunger but also prospectively reduce future hunger. To achieve this prospective reduction, it is necessary to collect information about external states of the world. This produces a form of 'epistemic value' that acts as a kind of uncertainty bonus driving actions towards unfamiliar states, much like the directed exploration strategy discussed earlier.

Active inference is a particular implementation of planning as inference, a family of algorithms that treat the policy as a latent variable, which is inferred conditional on the attainment of some goal state (for example, maximizing cumulative reward)[98]. This framework leads to a new interpretation of dopamine's role in reinforcement learning and decision making[17,97,99,100]. Instead of reporting RPEs, active inference models assert that dopamine reports the estimated precision (inverse variance) of the inferred policy. The precision corresponds to the agent's confidence that the policy it is currently following is optimal. At a neurobiological level, precision has been hypothesized to be implemented by gain modulation of neurons encoding the action policy, a function that is consistent with some prior computational models[96] and experimental data[101,102] on gain modulation, although little direct evidence exists for the role of gain modulation in the control of action policies. In effect, precision acts as an inverse temperature parameter, but it is now placed under the control of a continuously updated generative model, which implies that the policy stochasticity will change as beliefs are updated (policies will be more deterministic when beliefs are more precise). This theory is closely related to, and in some sense rationalizes, the random exploration strategy described earlier. In addition, the theory can also rationalize directed exploration strategies: uncertainty bonuses correspond to epistemic terms in the free energy that motivate actions to reduce future uncertainty[7,96].

Conclusions

Uncertainty plays a central role in modulating, and being modulated by, dopamine. A new generation of computational models have begun to formalize this interplay, accompanied by creative empirical tests of the theoretical predictions. We have shown how three different forms of uncertainty (associated with states, values and policies) affect the dopamine system in distinct, but computationally coherent, ways. State uncertainty affects the dopamine system via a probability distribution over states (the belief state), and values are defined as functions of the belief state. Value uncertainty affects the dopamine system via a probability distribution over the parameters of the value function. Finally, policy uncertainty affects the dopamine

Glossary

Active inference
The hypothesis that biological agents will take actions to reduce expected surprise.

Free-energy principle
The hypothesis that the objective of brain function is to minimize expected (average) surprise.

Posterior probability distribution
The conditional probability of latent variables (for example, hidden states) conditional on observed variables (for example, sensory data).

Sufficient statistic
A function of a data sample that completely summarizes the information contained in the data about the parameters of a probability distribution.

Value function
The mapping from states to long-term expected future rewards (typically discounted to reflect a preference for sooner over later rewards).

system via a probability distribution over the animal's actions. Under some accounts, dopamine levels may directly encode the precision (inverse variance) of the policy, thereby controlling the exploration–exploitation trade-off.

It is important to note that these different forms of uncertainty do not directly impinge on dopamine neurons. Rather, they enter into different parts of the information processing architecture in different ways. State uncertainty enters at the level of inputs to the striatum (putatively from the mPFC). Weight uncertainty enters at the level of striatal synapses. Policy uncertainty enters at the level of striatal outputs and possibly other areas. It is also important to note that downstream areas may receive information about uncertainty that does not rely on the dopamine system.

We see several important future directions. First, our treatment of state uncertainty assumed that the state space is known but partially observable. However, in reality, animals may also have uncertainty about the state space itself. This poses a structure learning problem for the brain. Although models have been developed to explain how structure learning might explain a range of reinforcement learning phenomena[103], we still lack a plausible neurobiological implementation. One speculative role for dopamine is to drive structure learning through RPEs. Some theories posit that sufficiently large RPEs will not lead to updating values but, rather, to updating the structural representation[104,105]; however, currently, there is no direct evidence for such a role for dopamine.

Second, we know very little about the hypothetical basis functions represented in the striatum. Although models for specific state spaces (for example, temporally defined states) have received some support[45], we still lack a general theory and adequate experimental tests. This question could be attacked using a combination of model-based and data-driven techniques; for example, by parameterizing a flexible space of basis functions and then fitting an encoding model for this space to striatal ensemble activity.

Third, the belief state framework addresses only some of the problems facing the standard TD model. A number of experiments have documented dopamine responses to non-rewarding stimuli that might be designated as 'sensory prediction errors'. Some of these findings can be accommodated by broadening the conceptualization of error that dopamine neuron activity encodes[106]. This broadened perspective is not intrinsically opposed to the belief state framework, and future work could fruitfully bridge these perspectives. In particular, Gardner et al.[106] have proposed that dopamine signalling reports a generalized prediction error defined over a collection of predictive features known as the successor representation[107]. Each predictive feature encodes an expectation of how often a particular sensory cue will be encountered soon, and these expectations can be updated using a form of TD learning with generalized prediction errors defined over these features. RPEs are a special case of such generalized prediction errors applied to a value (a cumulative reward) feature, and hence the framework is broadly compatible with the classical TD interpretation of dopamine as reporting RPEs. The successor representation does not, however, encode uncertainty about the predictive features. The Kalman TD model can in principle capture such uncertainty by generalizing the notion of value uncertainty to other predictive features.

Fourth, the precision account of tonic dopamine seems ill-equipped to explain the role of tonic dopamine in cognitive and physical effort[108–110]. One influential account of this role is the idea that tonic dopamine invigorates action through the encoding of average reward[111]. It is an open question how to adequately reconcile the average reward and precision accounts.

In closing, we note that progress in our understanding of belief state computations has been driven largely by theory-driven experiments. The theories described here make strong predictions that are highly unlikely under alternative accounts. We see this approach as a paradigmatic example of how computational models can be put to work in the service of experimental research, and vice versa.

Samuel J. Gershman[1]* *and Naoshige Uchida*[2]

[1]*Department of Psychology, Center for Brain Science, Harvard University, Cambridge, MA, USA.*

[2]*Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University, Cambridge, MA, USA.*

**e-mail: gershman@fas.harvard.edu*

https://doi.org/10.1038/s41583-019-0220-7

Published online 30 September 2019

1. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* **40**, 373–394 (2017).
2. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
3. Courville, A. C., Daw, N. D. & Touretzky, D. S. Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.* **10**, 294–300 (2006).
4. Gershman, S. J., Blei, D. M. & Niv, Y. Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
5. Gershman, S. J. A Unifying probabilistic view of associative learning. *PLOS Comput. Biol.* **11**, e1004567 (2015).
6. Kakade, S. & Dayan, P. Acquisition and extinction in autoshaping. *Psychol. Rev.* **109**, 533–544 (2002).
7. Friston, K. et al. Active inference and epistemic value. *Cogn. Neurosci.* **6**, 187–214 (2015).
8. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173**, 34–42 (2018).
9. Speekenbrink, M. & Konstantinidis, E. Uncertainty and exploration in a restless bandit problem. *Top. Cogn. Sci.* **7**, 351–367 (2015).
10. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
11. Ma, W. J. & Jazayeri, M. Neural coding of uncertainty and probability. *Annu. Rev. Neurosci.* **37**, 205–220 (2014).
12. Rao, R. P. N. Decision making under uncertainty: a neural model based on partially observable Markov decision processes. *Front. Comput. Neurosci.* **4**, 146 (2010).
13. Daw, N. D., Courville, A. C. & Touretzky, D. S. Representation and timing in theories of the dopamine system. *Neural Comput.* **18**, 1637–1677 (2006).
14. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
15. Jazayeri, M. & Movshon, J. A. Optimal representation of sensory information by neural populations. *Nat. Neurosci.* **9**, 690–696 (2006).
16. Grabska-Barwińska, A. et al. A probabilistic approach to demixing odors. *Nat. Neurosci.* **20**, 98–106 (2017).
17. Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P. & Pezzulo, G. Active inference: a process theory. *Neural Comput.* **29**, 1–49 (2017).
18. Buesing, L., Bill, J., Nessler, B. & Maass, W. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLOS Comput. Biol.* **7**, e1002211 (2011).
19. Pecevski, D., Buesing, L. & Maass, W. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons. *PLOS Comput. Biol.* **7**, e1002294 (2011).
20. Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–660 (2016).
21. Orbán, G., Berkes, P., Fiser, J. & Lengyel, M. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* **92**, 530–543 (2016).
22. Ting, C.-C., Yu, C.-C., Maloney, L. T. & Wu, S.-W. Neural mechanisms for integrating prior knowledge and likelihood in value-based probabilistic inference. *J. Neurosci.* **35**, 1792–1805 (2015).
23. Yoshida, W. & Ishii, S. Resolution of uncertainty in prefrontal cortex. *Neuron* **50**, 781–789 (2006).
24. Yoshida, W., Seymour, B., Friston, K. J. & Dolan, R. J. Neural mechanisms of belief inference during cooperative games. *J. Neurosci.* **30**, 10744–10751 (2010).
25. Fleming, S. M., van der Putten, E. J. & Daw, N. D. Neural mediators of changes of mind about perceptual decisions. *Nat. Neurosci.* **21**, 617–624 (2018).
26. Kumaran, D., Banino, A., Blundell, C., Hassabis, D. & Dayan, P. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron* **92**, 1135–1147 (2016).
27. Turner, M. S., Cipolotti, L., Yousry, T. A. & Shallice, T. Confabulation: damage to a specific inferior medial prefrontal system. *Cortex* **44**, 637–648 (2008).
28. Karlsson, M. P., Tervo, D. G. R. & Karpova, A. Y. Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science* **338**, 135–139 (2012).
29. Fuhs, M. C. & Touretzky, D. S. Context learning in the rodent hippocampus. *Neural Comput.* **19**, 3173–3215 (2007).
30. Dufort, R. H., Guttman, N. & Kimble, G. A. One-trial discrimination reversal in the white rat. *J. Comp. Physiol. Psychol.* **47**, 248–249 (1954).
31. Pubols, B. H. Jr. Serial reversal learning as a function of the number of trials per reversal. *J. Comp. Physiol. Psychol.* **55**, 66–68 (1962).

32. Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.* **104**, 1068–1076 (2010).
33. Gallistel, C. R., Mark, T. A., King, A. P. & Latham, P. E. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *J. Exp. Psychol. Anim. Behav. Process.* **27**, 354–372 (2001).
34. Jang, A. I. et al. The role of frontal cortical and medial-temporal lobe brain areas in learning a Bayesian prior belief on reversals. *J. Neurosci.* **35**, 11751–11760 (2015).
35. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–8367 (2006).
36. Mondragón, E., Alonso, E. & Kokkola, N. Associative learning should go deep. *Trends Cogn. Sci.* **21**, 822–825 (2017).
37. Gibbon, J. Scalar expectancy theory and weber's law in animal timing. *Psychol. Rev.* **84**, 279–325 (1977).
38. Gibbon, J., Church, R. M. & Meck, W. H. Scalar timing in memory. *Ann. NY Acad. Sci.* **423**, 52–77 (1984).
39. Shi, Z., Church, R. M. & Meck, W. H. Bayesian optimization of time perception. *Trends Cogn. Sci.* **17**, 556–564 (2013).
40. Petter, E. A., Gershman, S. J. & Meck, W. H. Integrating models of interval timing and reinforcement learning. *Trends Cogn. Sci.* **22**, 911–922 (2018).
41. Ma, W. J., Beck, J. M., Latham, P. E. & Pouget, A. Bayesian inference with probabilistic population codes. *Nat. Neurosci.* **9**, 1432–1438 (2006).
42. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Comput.* **20**, 3034–3054 (2008).
43. Ludvig, E. A., Sutton, R. S. & Kehoe, E. J. Evaluating the TD model of classical conditioning. *Learn. Behav.* **40**, 305–319 (2012).
44. Gershman, S. J., Moustafa, A. A. & Ludvig, E. A. Time representation in reinforcement learning models of the basal ganglia. *Front. Comput. Neurosci.* **7**, 194 (2014).
45. Mello, G. B. M., Soares, S. & Paton, J. J. A scalable population code for time in the striatum. *Curr. Biol.* **25**, 1113–1122 (2015).
46. Akhlaghpour, H. et al. Dissociated sequential activity and stimulus encoding in the dorsomedial striatum during spatial working memory. *eLife* **5**, e19507 (2016).
47. Kim, J., Kim, D. & Jung, M. W. Distinct dynamics of striatal and prefrontal neural activity during temporal discrimination. *Front. Integr. Neurosci.* **12**, 34 (2018).
48. Bakhurin, K. I. et al. Differential encoding of time by prefrontal and striatal network dynamics. *J. Neurosci.* **37**, 854–870 (2017).
49. Adler, A. et al. Temporal convergence of dynamic cell assemblies in the striato-pallidal network. *J. Neurosci.* **32**, 2473–2484 (2012).
50. Emmons, E. B. et al. Rodent medial frontal control of temporal processing in the dorsomedial striatum. *J. Neurosci.* **37**, 8718–8733 (2017).
51. Gouvêa, T. S. et al. Striatal dynamics explain duration judgments. *eLife* **4**, e11386 (2015).
52. Takahashi, Y. K., Langdon, A. J., Niv, Y. & Schoenbaum, G. Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat VTA depends on ventral striatum. *Neuron* **91**, 182–193 (2016).
53. Wiener, S. I. Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. *J. Neurosci.* **13**, 3802–3817 (1993).
54. Lavoie, A. M. & Mizumori, S. J. Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. *Brain Res.* **638**, 157–168 (1994).
55. Caan, W., Perrett, D. I. & Rolls, E. T. Responses of striatal neurons in the behaving monkey. 2. Visual processing in the caudal neostriatum. *Brain Res.* **290**, 53–65 (1984).
56. Brown, V. J., Desimone, R. & Mishkin, M. Responses of cells in the tail of the caudate nucleus during visual discrimination learning. *J. Neurophysiol.* **74**, 1083–1094 (1995).
57. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).
58. Schultz, W. & Romo, R. Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *J. Neurophysiol.* **63**, 607–624 (1990).
59. Kobayashi, S. & Schultz, W. Reward contexts extend dopamine signals to unrewarded stimuli. *Curr. Biol.* **24**, 56–62 (2014).
60. Matsumoto, H., Tian, J., Uchida, N. & Watabe-Uchida, M. Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. *eLife* **5**, e17328 (2016).
61. Hollerman, J. R. & Schultz, W. Dopamine neurons report an error in the temporal prediction of reward during learning. *Nat. Neurosci.* **1**, 304–309 (1998).
62. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589 (2017).
63. Fiorillo, C. D., Newsome, W. T. & Schultz, W. The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.* **11**, 966–973 (2008).
64. Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y. & Hikosaka, O. Dopamine neurons can represent context-dependent prediction error. *Neuron* **41**, 269–280 (2004).
65. Starkweather, C. K., Gershman, S. J. & Uchida, N. The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron* **98**, 616–629.e6 (2018).
66. Babayan, B. M., Uchida, N. & Gershman, S. J. Belief state representation in the dopamine system. *Nat. Commun.* **9**, 1891 (2018).
67. Nomoto, K., Schultz, W., Watanabe, T. & Sakagami, M. Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. *J. Neurosci.* **30**, 10692–10702 (2010).
68. Lak, A., Nomoto, K., Keramati, M., Sakagami, M. & Kepecs, A. Midbrain dopamine neurons signal belief in choice accuracy during a perceptual decision. *Curr. Biol.* **27**, 821–832 (2017).
69. Sarno, S., de Lafuente, V., Romo, R. & Parga, N. Dopamine reward prediction error signal codes the temporal evaluation of a perceptual decision report. *Proc. Natl Acad. Sci. USA* **114**, E10494–E10503 (2017).
70. Ghavamzadeh, M., Mannor, S., Pineau, J. & Tamar, A. *Bayesian Reinforcement Learning: A Survey* (Now Publishers, 2015).
71. Gershman, S. J. Dopamine, inference, and uncertainty. *Neural Comput.* **29**, 3311–3326 (2017).
72. Kamin, L. J. in *Punishment and Aversive Behavior* (eds Campbell, B. A. & Church, R. M.) 279–296 (Appleton-Century-Crofts, 1969).
73. Rescorla, R. A. & Wagner, A. R. in *Classical Conditioning II: Recent Research and Theory* (eds Black, A. H. & Prokasy, W. F.) 64–99 (Appleton-Century-Crofts, 1972).
74. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412**, 43–48 (2001).
75. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
76. Miller, R. R. & Matute, H. Biological significance in forward and backward blocking: resolution of a discrepancy between animal conditioning and human causal judgment. *J. Exp. Psychol. Gen.* **125**, 370–386 (1996).
77. Urushihara, K. & Miller, R. R. Backward blocking in first-order conditioning. *PsycEXTRA Dataset* https://doi.org/10.1037/e527342012-212 (2007).
78. Blaisdell, A. P., Gunther, L. M. & Miller, R. R. Recovery from blocking achieved by extinguishing the blocking CS. *Anim. Learn. Behav.* **27**, 63–76 (1999).
79. Dayan, P. & Kakade, S. Explaining away in weight space. *Adv. Neural Inf. Process. Syst.* **13**, 451–457 (2001).
80. Miller, R. R. & Witnauer, J. E. Retrospective revaluation: the phenomenon and its theoretical implications. *Behav. Process.* **123**, 15–25 (2016).
81. Lubow, R. E. Latent inhibition. *Psychol. Bull.* **79**, 398–407 (1973).
82. Aguado, L., Symonds, M. & Hall, G. Interval between preexposure and test determines the magnitude of latent inhibition: Implications for an interference account. *Anim. Learn. Behav.* **22**, 188–194 (1994).
83. Sadacca, B. F., Jones, J. L. & Schoenbaum, G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife* **5**, e13665 (2016).
84. Young, A. M., Joseph, M. H. & Gray, J. A. Latent inhibition of conditioned dopamine release in rat nucleus accumbens. *Neuroscience* **54**, 5–9 (1993).
85. Frank, M. J. & Claus, E. D. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychol. Rev.* **113**, 300–326 (2006).
86. Deco, G. & Rolls, E. T. Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. *Cereb. Cortex* **15**, 15–30 (2005).
87. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
88. Sadacca, B. F. et al. Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *eLife* **7**, e30373 (2018).
89. Jones, J. L. et al. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338**, 953–956 (2012).
90. Payzan-LeNestour, É. & Bossaerts, P. Do not bet on the unknown versus try to find out more: estimation uncertainty and 'unexpected uncertainty' both modulate exploration. *Front. Neurosci.* **6**, 150 (2012).
91. Schulz, E., Konstantinidis, E. & Speekenbrink, M. Putting bandits into context: How function learning supports decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 927–943 (2018).
92. Myers, J. L. & Sadler, E. Effects of range of payoffs as a variable in risk taking. *J. Exp. Psychol.* **60**, 306–309 (1960).
93. Busemeyer, J. R. & Townsend, J. T. Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. *Psychol. Rev.* **100**, 432–459 (1993).
94. Gershman, S. J. Uncertainty and exploration. *Decision* **6**, 277–286 (2019).
95. Frank, M. J., Doll, B. B., Oas-Terpstra, J. & Moreno, F. Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nat. Neurosci.* **12**, 1062–1068 (2009).
96. Humphries, M. D., Khamassi, M. & Gurney, K. Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Front. Neurosci.* **6**, 9 (2012).
97. Pezzulo, G., Rigoli, F. & Friston, K. J. Hierarchical active inference: a theory of motivated control. *Trends Cogn. Sci.* **22**, 294–306 (2018).
98. Botvinick, M. & Toussaint, M. Planning as inference. *Trends Cogn. Sci.* **16**, 485–488 (2012).
99. FitzGerald, T. H. B., Dolan, R. J. & Friston, K. Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.* **9**, 136 (2015).
100. Friston, K. J. et al. Dopamine, affordance and active inference. *PLOS Comput. Biol.* **8**, e1002327 (2012).
101. Weele, C. M. V. et al. Dopamine enhances signal-to-noise ratio in cortical-brainstem encoding of aversive stimuli. *Nature* **563**, 397–401 (2018).
102. Thurley, K., Senn, W. & Lüscher, H.-R. Dopamine increases the gain of the input-output response of rat prefrontal pyramidal neurons. *J. Neurophysiol.* **99**, 2985–2997 (2008).
103. Gershman, S. J., Norman, K. A. & Niv, Y. Discovering latent causes in reinforcement learning. *Curr. Opin. Behav. Sci.* **5**, 43–50 (2015).
104. Gershman, S. J., Monfils, M.-H., Norman, K. A. & Niv, Y. The computational nature of memory modification. *eLife* **6**, e23763 (2017).
105. Redish, A. D., Jensen, S., Johnson, A. & Kurth-Nelson, Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychol. Rev.* **114**, 784–805 (2007).
106. Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc. Biol. Sci.* **285**, 20181645 (2018).
107. Gershman, S. J. The successor representation: its computational logic and neural substrates. *J. Neurosci.* **38**, 7193–7200 (2018).
108. Le Bouc, R. et al. Computational dissection of dopamine motor and motivational functions in humans. *J. Neurosci.* **36**, 6623–6633 (2016).
109. Walton, M. E. & Bouret, S. What is the relationship between dopamine and effort? *Trends Neurosci.* **42**, 79–91 (2019).

110. Westbrook, A. & Braver, T. S. Dopamine does double duty in motivating cognitive effort. *Neuron* **91**, 708 (2016).
111. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507–520 (2007).
112. Sutton, R. S. Learning to predict by the methods of temporal differences. *Mach. Learn.* **3**, 9–44 (1988).
113. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134 (1998).
114. Pan, W.-X., Schmidt, R., Wickens, J. R. & Hyland, B. I. Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. *J. Neurosci.* **25**, 6235–6242 (2005).
115. Menegas, W., Babayan, B. M., Uchida, N. & Watabe-Uchida, M. Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife* **6**, e21886 (2017).
116. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive coding of reward value by dopamine neurons. *Science* **307**, 1642–1645 (2005).
117. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).

Author contributions

The authors contributed equally to all aspects of the article.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information

Nature Reviews Neuroscience thanks J. Pearson and the other, anonymous, reviewers for their contribution to the peer review of this work.