Published in final edited form as:

Nat Rev Neurosci. 2019 November ; 20(11): 703–714. doi:10.1038/s41583-019-0220-7.

Believing in Dopamine

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

[1]Department of Psychology, Center for Brain Science, Harvard University

[2]Department of Molecular and Cellular Biology, Center for Brain Science, Harvard University

Abstract

Dopamine signals are widely thought to report reward prediction errors that drive learning in the basal ganglia. However, dopamine has also been implicated in a variety of probabilistic computations, such as encoding uncertainty and controlling exploration. These different facets of dopamine can be brought together under a common reinforcement learning framework. The key idea is that multiple sources of uncertainty impinge upon 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/basal ganglia circuit for reinforcement learning, and are 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.

Introduction

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 also appears to participate in a variety of probabilistic computations, including the encoding of uncertainty and the control of uncertainty-guided exploration. The purpose of this review is to bring together these different strands into a common reinforcement learning framework.

The key ingredients of reinforcement learning theories are (1) state, (2) value, and (3) policy. In reinforcement learning, values are computed based on the state of world that the animal currently occupies. A state is collectively defined by the animal's location, 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 different forms of uncertainty encompassing all of these ingredients – state, value and policy.

Competing interests statement

The authors have no competing interests to declare.

First, animals commonly do not have full information about which state they are currently occupying. Rather, they receive sensory data that provides ambiguous information about the current state.[3,4] For example, an animal might sense an odor plume to infer the hidden location of a food source. Because many different locations could be compatible with the odor plume to varying degrees, the normatively correct strategy is to compute the ***posterior probability distribution*** of the food location conditional on the odor information. This computation is stipulated by Bayes' rule: P(location|odor) is the product of the likelihood P(odor|location) and the prior P(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 (e.g., the value of a patch is approximated by a weighted sum of its features such as size and resource density). Because these parameters are unknown, the animal has uncertainty about them that is gradually resolved through experience of rewards in different states. Whereas standard models of learning, such as the temporal difference (TD) model, update point estimates of the parameters (i.e., a single set of parameter values), other models encode uncertainty about the parameters in the form of a probability distribution over 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 tend to be selected that take the animal to rewarding states. 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 in order 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 dopamine activity, by impinging on different stages of the information processing architecture for reinforcement learning (Fig 1). As we elaborate below, these effects can be formalized in terms of Bayesian reinforcement learning principles. The Bayesian framework significantly enriches the traditional RPE interpretation of dopamine, 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 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 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 (lion vs. no lion), the normatively correct representation of uncertainty is the posterior 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 distributions.[11] From an RL perspective, the question is how the animal should use the posterior 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 above, 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 RL problem is what enables efficient algorithms, such as dynamic programming (value iteration) and temporal difference (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 distribution encodes all the available information for predicting future reward. The posterior 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 as the "belief state."

The belief state plays the same role as other state representations in the standard RL 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. We will unpack each step of this machinery as it applies to belief states. First, we describe how belief states are computed in the medial prefrontal cortex (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.

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.[17] One limitation of such a "labeled 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,[18,19] or by approximately sampling from the posterior to construct a nonparametric approximation.[20-23]

Several lines of evidence point to the mPFC as a candidate locus for belief state computation. Changes in mPFC activity track updating of the posterior distribution[24-28], and damage to the mPFC is associated with aberrant belief formation, such as confabulation.[29] Changes in mPFC activity ("network resets") are also associated with the onset of behavioral variability,[30] consistent with the idea that the variance of the posterior distribution (representing the animal's uncertainty) controls the randomness of the action policy, as discussed further below.[8,30]

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 behavior,[31-34] a phenomenon inconsistent with models of learning in which reward predictions are relearned after each reversal. As several authors have noted,[31,35] reversal learning is better modeled 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 mPFC appear to leave rapid reacquisition intact, but increase the rate of reversal, consistent with either a reduced evidence threshold or an increased estimate of the reversal probability.[36] A functional MRI study of reversal learning in humans[37] found that mPFC activity is sensitive to hidden state inference.

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

Value function approximation.

From the RL perspective, the goal of 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 more a 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 more complex nonlinear value function approximation is possible,[38] 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 state (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.[39] Pacemaker-accumulator models[40] explain this phenomenon mechanistically: an accumulator noisily counts pulses emitted from a pacemaker, and these counts are compared to a reference retrieved from memory, corrupted by multiplicative noise. From a Bayesian perspective,[41,42] 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 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 labeled 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[43] (i.e., for longer intervals).

If we assume that striatal basis function neurons receive input from a sub-population of similarly tuned cortical belief state neurons, then the temporal profile of striatal activation will be more smeared out for longer intervals, due 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,[44-46] 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[47-50] and primate[51] striatum, whereas other studies have reported monotonic tuning (i.e., ramping).[52] Consistent with a causal role for striatal time cells in downstream computations, the temporal specificity of both behavior[53] and dopamine activity[54] depends on the integrity of the striatum.

In sum, data from interval timing experiments is 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[55,56] or visually[57,58] 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 assuming a fully observable state.[12,13,59] For example, in one study[60] 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, finding that dopamine responses generalize to unrewarded stimuli when they occurred in the same context as rewarded stimuli,[61] and more recently it was found that this reward generalization can apply even to aversive stimuli.[62] One possible explanation is that the monkey was initially uncertain about which box was opening, and therefore its value estimate defined on the belief state would reflect a mixture of the two box-specific values, producing a positive RPE.[2,59] 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 0 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, or ISI). The delay between the reward and the onset of the next cue is the intertrial interval (ITI). Daw and colleagues[13] modeled 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.

Under this 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. In 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 0. This prediction is consistent with empirical observations: no suppression of dopamine activity is observed at the expected time of reward.[63]

Recent studies have built on these findings, pursuing a more detailed empirical test of the model's 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.[64,65] 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 the RPE will decrease. When rewards are delivered stochastically (10% of rewards are omitted), the pattern changes radically: dopamine neurons respond *more* 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 go down and the RPE to go up. Several additional analyses also ruled out an alternative account based on subjective hazard functions.[65,66]

Consistent with the hypothesized role for mPFC in state inference, the effect of state uncertainty on dopamine responses is disrupted by inactivation of mPFC.[64,67] Specifically, the monotonic increase in dopamine response as a function of ISI in the 90% reward condition is flattened out under mPFC inactivation. Importantly, there is no effect of mPFC inactivation on the monotonic decrease in the 100% reward condition. Furthermore, the sensitivity of other dopamine responses to interval timing (e.g. a 'dip' during reward omission) remained intact. These results suggested that 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 (Fig 2B).[68] Mice first alternated between two conditions distinguished only by the reward magnitude. On small reward blocks, animals received an odor cue and then shortly afterward a small water reward. Large reward blocks were identical (including the same odor 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 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.

Dopamine responses conformed to this predicted "zig-zag" pattern. The same pattern was also reflected in anticipatory licking behavior: 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 behavior). The standard TD model, in contrast, can only predict a monotonic pattern of dopamine responses (no zig-zag) when using a single hidden state for all blocks, and could not as accurately predict variations in anticipatory licking.

Belief-dependent reward prediction errors in perceptual decision making.

Another line of evidence for the belief-dependence of dopamine comes from perceptual decision making tasks. In the most extensively analyzed study, dopamine neurons were recorded while monkeys performed a random dot motion discrimination task.[68,69] On each trial, the monkeys saw a set of moving dots, with some proportion of dots (the coherence) moving either left or right. The monkeys reported perceived direction by saccading to one of two targets. A key finding from this study was that the size of the stimulus-evoked dopamine 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 re-analysis of these data further verified critical predictions of the belief state model.[70] Conventional TD models reflect stimulus-reward associations as mentioned above, and predict that the stimulus-evoked response should not be modulated by the animal's choice. In contrast, a belief-state TD model uses the inferred stimulus, which drives animal's choice and, at the same time, modulates the animal's confidence about receiving reward. Supporting the belief state TD model, their analysis showed that the dopamine response depends jointly on performance (correct vs. incorrect) as well as coherence (Fig 3).

Related results have been found using a vibrotactile detection task.[71] Monkeys judged whether or not a weak vibrotactile stimulus occurred during an observation period. On hit trials (monkeys correctly detected the stimulus), the dopamine response to the stimulus increases monotonically with stimulus amplitude, which can be thought of as analogous to coherence in terms of its effect on the belief state. Furthermore, the dopamine response is higher on false alarm trials (monkeys incorrectly reported the stimulus when it did not occur) than on correct rejection trials (monkeys correctly reported that the stimulus did not occur). This indicates that dopamine responses reflect subjective beliefs about the stimulus rather than the objective stimulus, consistent with the belief state model.

Value uncertainty

In the models described above, 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.[74] One analytically tractable and neurally plausible special case is "Kalman temporal difference 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.[75]

Behavioral evidence for value uncertainty.

One of the classic pieces of evidence for error-driven learning comes from "blocking" experiments in which an animal first learns to associate a stimulus with reward (A+) and then the stimulus is paired with another stimulus (B) while continuing to be rewarded (AB +). In the final phase, the animal is tested on the second stimulus without reward (B−). Despite the fact that B was reliably paired with reward, the test phase typically reveals a weak or absent conditioned response; evidently, the association between A and reward "blocks" learning about the association between B and reward.[75,76] This indicates that correlation between a stimulus and reward is not a sufficient condition for learning. The Rescorla-Wagner model[77] 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. 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.[77,78] Moreover, blocking can be counteracted by optogenetic stimulation of dopamine neurons during compound training.[79]

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 (so-called *backward blocking*, to contrast it with the *forward blocking* effect described in the last paragraph).[79-81] 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 B-reward association to be modified, a process that is not allowed under the Rescorla-Wagner model or the standard TD model; in these models, errors can only drive learning of present stimuli. An answer to this problem is provided by a Bayesian treatment of the TD model, known as Kalman TD learning[5,82] (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 AB cannot exceed the sum of 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 go down, thus producing backward blocking.[83] 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.[84] For example, in the forward blocking paradigm, extinguishing A following compound training increases responding to B in the test phase.[82] Another application is to single cue learning: when a neutral stimulus is pre-exposed (presented repeatedly without reward), subsequent conditioning of that stimulus is retarded, a phenomenon known as *latent inhibition*.[85] 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.[86] 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.[87] In the preconditioning phase, stimulus A is paired serially with stimulus B. Note that because no reward is 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.[75] In the second phase, B is paired with reward, and finally in the third phase the response to A is probed. Behaviorally, rats show a conditioned response to A, despite the fact that A was never paired with reward. This 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, using slow microdialysis methods in aversive conditioning, has provided some support for the Bayesian interpretation of latent inhibition described above. The stimulus-evoked dopamine response during the conditioning phase is reduced following pre-exposure,[88] consistent with the assumption that RPEs will propagate more slowly to the cue onset for the preexposed cue.[75] More experiments are needed to confirm the generality of these findings.

Stimulus transformation in orbitofrontal cortex underlying value uncertainty.

As described in Box 3, 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. This transformed representation can then be linearly mapped to reward predictions, and the posterior 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,[89,90] and has been implicated in state representation more broadly.[91] Consistent with this hypothesis, neurons in the orbitofrontal cortex come to reflect the associative structure of sensory preconditioning,[92] and lesions of orbitofrontal cortex impairs the sensory preconditioning effect.[91,93]

Policy uncertainty

Ultimately, the brain's RL 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.[8,9,96] 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.[8,97] A classic example of such "random exploration" is the payoff variability effect: choices are more variable when rewards are more variable.[98,99]

Recent studies have shown that these strategies can be simultaneously identified in human choice behavior,[8] and can be manipulated orthogonally.[100] 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 is more variable. 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 is greatest when both options are risky, compared to when both options are safe.

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 behavior using probit regression, we can extract directed and random exploration effects from the estimated coefficients.[8,100]

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.[101] Variation in DARPP-32, which primarily controls striatal dopamine levels, was selectively associated with random exploration, consistent with prior biophysical modeling[101,102] (although this modeling work did not directly simulate the effects of DARPP-32 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 fulfill the predictions of a generative model.[7,19] At first glance, this seems to be in direct opposition to the principle that actions should be taken to gain information about the world (e.g., 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.[103] 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 actions that will prospectively reduce future hunger. In order 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 a kind of uncertainty bonus driving actions towards unfamiliar states, much like the directed exploration strategy discussed above.

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 (e.g., maximizing cumulative reward).[104] This framework leads to a new interpretation of dopamine's role in reinforcement learning and decision making.[19,103,105,106] 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[102] and experimental data.[107,108] In effect, precision acts as an inverse temperature parameter, but now placed under 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 above. 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.[7,102]

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 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 medial prefrontal cortex). Weight uncertainty enters at the level of striatal synapses. Policy uncertainty enters at the level of striatal outputs and possibly other areas.

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,[109] we still lack a plausible neurobiological implementation. One speculative role for dopamine would be 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,[110,111] but 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 (e.g. temporally-defined states) have received some support,[47] 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 only addresses 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 encodes.[112] This broadened perspective is not intrinsically opposed to the belief state framework, and future work could fruitfully bridge these perspectives. In particular, Gardner and colleagues[112] have proposed that dopamine reports a generalized prediction error defined over a collection of predictive features known as the *successor representation*.[113] Each predictive feature encodes an expectation of how often a particular sensory cue will be encountered in the near future, and these expectations can be updated using a form of TD learning with generalized prediction errors defined over these features. Reward prediction errors are a special case of such generalized predictions errors applied to a value (cumulative reward) feature, and hence the framework is broadly compatible with the classical TD interpretation of dopamine as reporting reward prediction errors. 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.[114-116] One influential account of this role is the idea that tonic dopamine invigorates action through the encoding of average reward.[117] 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.

Glossary

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

Posterior probability distribution
the conditional probability of latent variables (e.g., hidden states) conditional on observed variables (e.g., 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

Free energy principle

the hypothesis that the objective of brain function is to minimize expected (average) surprise

Active inference

the hypothesis that biological agents will take actions to reduce expected surprise

References

1. Watabe-Uchida M, Eshel N & Uchida N Neural Circuitry of Reward Prediction Error. Annu. Rev. Neurosci 40, 373–394 (2017). [PubMed: 28441114]
2. Schultz W, Dayan P & Montague PR A neural substrate of prediction and reward. Science 275, 1593–1599 (1997). [PubMed: 9054347]
3. Courville AC, Daw ND & Touretzky DS Bayesian theories of conditioning in a changing world. Trends Cogn. Sci 10, 294–300 (2006). [PubMed: 16793323]
4. Gershman SJ, Blei DM & Niv Y Context, learning, and extinction. Psychol. Rev 117, 197–209 (2010). [PubMed: 20063968]
5. Gershman SJ A Unifying Probabilistic View of Associative Learning. PLoS Comput. Biol 11, e1004567 (2015). [PubMed: 26535896]
6. Kakade S & Dayan P Acquisition and extinction in autoshaping. Psychol. Rev 109, 533–544 (2002). [PubMed: 12088244]
7. Friston K et al. Active inference and epistemic value. Cogn. Neurosci 6, 187–214 (2015). [PubMed: 25689102]
8. Gershman SJ Deconstructing the human algorithms for exploration. Cognition 173, 34–42 (2018). [PubMed: 29289795]
9. Speekenbrink M & Konstantinidis E Uncertainty and exploration in a restless bandit problem. Top. Cogn. Sci 7, 351–367 (2015). [PubMed: 25899069]
10. Wilson RC, Geana A, White JM, Ludvig EA & Cohen JD Humans use directed and random exploration to solve the explore-exploit dilemma. J. Exp. Psychol. Gen 143, 2074–2081 (2014). [PubMed: 25347535]
11. Ma WJ & Jazayeri M Neural coding of uncertainty and probability. Annu. Rev. Neurosci 37, 205–220 (2014). [PubMed: 25032495]
12. Rao RPN Decision making under uncertainty: a neural model based on partially observable markov decision processes. Front. Comput. Neurosci 4, 146 (2010). [PubMed: 21152255]
13. Daw ND, Courville AC & Touretzky DS Representation and timing in theories of the dopamine system. Neural Comput. 18, 1637–1677 (2006). [PubMed: 16764517]
14. Gershman SJ & Daw ND Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. Annu. Rev. Psychol 68, 101–128 (2017). [PubMed: 27618944]
15. Sutton RS Learning to predict by the methods of temporal differences. Mach. Learn 3, 9–44 (1988).
16. Kaelbling LP, Littman ML & Cassandra AR Planning and acting in partially observable stochastic domains. Artif. Intell 101, 99–134 (1998).
17. Jazayeri M & Movshon JA Optimal representation of sensory information by neural populations. Nat. Neurosci 9, 690–696 (2006). [PubMed: 16617339]
18. Grabska-Barwińska A et al. A probabilistic approach to demixing odors. Nat. Neurosci 20, 98–106 (2017). [PubMed: 27918530]
19. Friston K, FitzGerald T, Rigoli F, Schwartenbeck P & Pezzulo G Active Inference: A Process Theory. Neural Comput. 29, 1–49 (2017). [PubMed: 27870614]

20. 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). [PubMed: 22096452]
21. 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). [PubMed: 22219717]
22. Haefner RM, Berkes P & Fiser J Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron 90, 649–660 (2016). [PubMed: 27146267]
23. Orbán G, Berkes P, Fiser J & Lengyel M Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. Neuron 92, 530–543 (2016). [PubMed: 27764674]
24. Ting C-C, Yu C-C, Maloney LT & Wu S-W Neural mechanisms for integrating prior knowledge and likelihood in value-based probabilistic inference. J. Neurosci 35, 1792–1805 (2015). [PubMed: 25632152]
25. Yoshida W & Ishii S Resolution of uncertainty in prefrontal cortex. Neuron 50, 781–789 (2006). [PubMed: 16731515]
26. Yoshida W, Seymour B, Friston KJ & Dolan RJ Neural mechanisms of belief inference during cooperative games. J. Neurosci 30, 10744–10751 (2010). [PubMed: 20702705]
27. Fleming SM, van der Putten EJ & Daw ND Neural mediators of changes of mind about perceptual decisions. Nat. Neurosci 21, 617–624 (2018). [PubMed: 29531361]
28. 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). [PubMed: 27930904]
29. Turner MS, Cipolotti L, Yousry TA & Shallice T Confabulation: damage to a specific inferior medial prefrontal system. Cortex 44, 637–648 (2008). [PubMed: 18472034]
30. Karlsson MP, Tervo DGR & Karpova AY Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. Science 338, 135–139 (2012). [PubMed: 23042898]
31. Fuhs MC & Touretzky DS Context learning in the rodent hippocampus. Neural Comput. 19, 3173–3215 (2007). [PubMed: 17970649]
32. Dufort RH, Guttman N & Kimble GA One-trial discrimination reversal in the white rat. J. Comp. Physiol. Psychol 47, 248–249 (1954). [PubMed: 13163264]
33. Pubols BH Jr. Serial reversal learning as a function of the number of trials per reversal. J. Comp. Physiol. Psychol 55, 66–68 (1962). [PubMed: 14489100]
34. Bromberg-Martin ES, Matsumoto M, Hong S & Hikosaka O A pallidus-habenula-dopamine pathway signals inferred stimulus values. J. Neurophysiol 104, 1068–1076 (2010). [PubMed: 20538770]
35. Gallistel CR, Mark TA, King AP & Latham PE 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). [PubMed: 11676086]
36. Jang AI 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). [PubMed: 26290251]
37. Hampton AN, Bossaerts P & O'Doherty JP The Role of the Ventromedial Prefrontal Cortex in Abstract State-Based Inference during Decision Making in Humans. Journal of Neuroscience 26, 8360–8367 (2006). [PubMed: 16899731]
38. Mondragón E, Alonso E & Kokkola N Associative Learning Should Go Deep. Trends Cogn. Sci 21, 822–825 (2017). [PubMed: 28668210]
39. Gibbon J Scalar expectancy theory and Weber's law in animal timing. Psychol. Rev 84, 279–325 (1977).
40. Gibbon J, Church RM & Meck WH Scalar timing in memory. Ann. N. Y. Acad. Sci 423, 52–77 (1984). [PubMed: 6588812]
41. Shi Z, Church RM & Meck WH Bayesian optimization of time perception. Trends Cogn. Sci 17, 556–564 (2013). [PubMed: 24139486]
42. Petter EA, Gershman SJ & Meck WH Integrating Models of Interval Timing and Reinforcement Learning. Trends Cogn. Sci 22, 911–922 (2018). [PubMed: 30266150]

43. Ma WJ, Beck JM, Latham PE & Pouget A Bayesian inference with probabilistic population codes. Nat. Neurosci 9, 1432–1438 (2006). [PubMed: 17057707]
44. Ludvig EA, Sutton RS & Kehoe EJ Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. Neural Comput. 20, 3034–3054 (2008). [PubMed: 18624657]
45. Ludvig EA, Sutton RS & Kehoe EJ Evaluating the TD model of classical conditioning. Learn. Behav 40, 305–319 (2012). [PubMed: 22927003]
46. Gershman SJ, Moustafa AA & Ludvig EA Time representation in reinforcement learning models of the basal ganglia. Front. Comput. Neurosci 7, 194 (2014). [PubMed: 24409138]
47. Mello GBM, Soares S & Paton JJ A scalable population code for time in the striatum. Curr. Biol 25, 1113–1122 (2015). [PubMed: 25913405]
48. Akhlaghpour H et al. Dissociated sequential activity and stimulus encoding in the dorsomedial striatum during spatial working memory. Elife 5, (2016).
49. Kim J, Kim D & Jung MW Distinct Dynamics of Striatal and Prefrontal Neural Activity During Temporal Discrimination. Front. Integr. Neurosci 12, (2018).
50. Bakhurin KI et al. Differential Encoding of Time by Prefrontal and Striatal Network Dynamics. J. Neurosci 37, 854–870 (2017). [PubMed: 28123021]
51. Adler A et al. Temporal convergence of dynamic cell assemblies in the striato-pallidal network. J. Neurosci 32, 2473–2484 (2012). [PubMed: 22396421]
52. Emmons EB et al. Rodent Medial Frontal Control of Temporal Processing in the Dorsomedial Striatum. J. Neurosci 37, 8718–8733 (2017). [PubMed: 28821670]
53. Gouvêa TS et al. Striatal dynamics explain duration judgments. Elife 4, (2015).
54. Takahashi YK, Langdon AJ, 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). [PubMed: 27292535]
55. Wiener SI Spatial and behavioral correlates of striatal neurons in rats performing a self-initiated navigation task. J. Neurosci 13, 3802–3817 (1993). [PubMed: 8366346]
56. Lavoie AM & Mizumori SJ Spatial, movement- and reward-sensitive discharge by medial ventral striatum neurons of rats. Brain Res. 638, 157–168 (1994). [PubMed: 8199856]
57. Caan W, Perrett DI & Rolls ET Responses of striatal neurons in the behaving monkey. 2. Visual processing in the caudal neostriatum. Brain Res. 290, 53–65 (1984). [PubMed: 6692139]
58. Brown VJ, Desimone R & Mishkin M Responses of cells in the tail of the caudate nucleus during visual discrimination learning. J. Neurophysiol 74, 1083–1094 (1995). [PubMed: 7500134]
59. Kakade S & Dayan P Dopamine: generalization and bonuses. Neural Netw. 15, 549–559 (2002). [PubMed: 12371511]
60. 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). [PubMed: 2329364]
61. Kobayashi S & Schultz W Reward contexts extend dopamine signals to unrewarded stimuli. Curr. Biol 24, 56–62 (2014). [PubMed: 24332545]
62. Matsumoto H, Tian J, Uchida N & Watabe-Uchida M Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. Elife 5, (2016).
63. Hollerman JR & Schultz W Dopamine neurons report an error in the temporal prediction of reward during learning. Nat. Neurosci 1, 304–309 (1998). [PubMed: 10195164]
64. Starkweather CK, Babayan BM, Uchida N & Gershman SJ Dopamine reward prediction errors reflect hidden-state inference across time. Nat. Neurosci 20, 581–589 (2017). [PubMed: 28263301]
65. Fiorillo CD, Newsome WT & Schultz W The temporal precision of reward prediction in dopamine neurons. Nat. Neurosci 11, 966–973 (2008). [PubMed: 18660807]
66. Nakahara H, Itoh H, Kawagoe R, Takikawa Y & Hikosaka O Dopamine neurons can represent context-dependent prediction error. Neuron 41, 269–280 (2004). [PubMed: 14741107]

67. Starkweather CK, Gershman SJ & Uchida N The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. Neuron 98, 616–629.e6 (2018). [PubMed: 29656872]
68. Babayan BM, Uchida N & Gershman SJ Belief state representation in the dopamine system. Nat. Commun 9, 1891 (2018). [PubMed: 29760401]
69. Nomoto K, Schultz W, Watanabe T & Sakagami M Temporally extended dopamine responses to perceptually demanding reward-predictive stimuli. J. Neurosci 30, 10692–10702 (2010). [PubMed: 20702700]
70. 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). [PubMed: 28285994]
71. 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. U. S. A 114, E10494–E10503 (2017). [PubMed: 29133424]
72. Pan W-X, Schmidt R, Wickens JR & Hyland BI Dopamine cells respond to predicted events during classical conditioning: evidence for eligibility traces in the reward-learning network. J. Neurosci 25, 6235–6242 (2005). [PubMed: 15987953]
73. Menegas W, Babayan BM, Uchida N & Watabe-Uchida M Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. Elife 6, (2017).
74. Ghavamzadeh M, Mannor S, Pineau J & Tamar A Bayesian Reinforcement Learning: A Survey. (Now Publishers, 2015).
75. Gershman SJ Dopamine, Inference, and Uncertainty. Neural Comput. 29, 3311–3326 (2017). [PubMed: 28957023]
76. Kamin LJ Predictability, surprise, attention, and conditioning in Punishment and Aversive Behavior (ed. Campbell BA and Church RM) 279–296 (Appleton-Century-Crofts, 1969).
77. Rescorla RA & Wagner AR A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement in Classical Conditioning II: Recent Research and Theory (ed. Black AH and Prokasy WF) 64–99 (Appleton-Century Crofts, 1972).
78. Waelti P, Dickinson A & Schultz W Dopamine responses comply with basic assumptions of formal learning theory. Nature 412, 43–48 (2001). [PubMed: 11452299]
79. Steinberg EE et al. A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci 16, 966–973 (2013). [PubMed: 23708143]
80. Miller RR & 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). [PubMed: 8945788]
81. Urushihara K & Miller RR Backward Blocking in First-Order Conditioning. PsycEXTRA Dataset (2007). doi:10.1037/e527342012-212
82. Blaisdell AP, Gunther LM & Miller RR Recovery from blocking achieved by extinguishing the blocking CS. Anim. Learn. Behav 27, 63–76 (1999).
83. Dayan P & Kakade S Explaining away in weight space. Adv. Neural Inf. Process. Syst 13, 451–457 (2001).
84. Miller RR & Witnauer JE Retrospective revaluation: The phenomenon and its theoretical implications. Behav. Processes 123, 15–25 (2016). [PubMed: 26342855]
85. Lubow RE Latent inhibition. Psychol. Bull 79, 398–407 (1973). [PubMed: 4575029]
86. 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).
87. Sadacca BF, Jones JL & Schoenbaum G Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. Elife 5, (2016).
88. Young AM, Joseph MH & Gray JA Latent inhibition of conditioned dopamine release in rat nucleus accumbens. Neuroscience 54, 5–9 (1993). [PubMed: 8515846]
89. Frank MJ & Claus ED Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol. Rev 113, 300–326 (2006). [PubMed: 16637763]

90. Deco G & Rolls ET Synaptic and spiking dynamics underlying reward reversal in the orbitofrontal cortex. Cereb. Cortex 15, 15–30 (2005). [PubMed: 15238449]
91. Wilson RC, Takahashi YK, Schoenbaum G & Niv Y Orbitofrontal cortex as a cognitive map of task space. Neuron 81, 267–279 (2014). [PubMed: 24462094]
92. Sadacca BF et al. Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. Elife 7, (2018).
93. Jones JL et al. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science 338, 953–956 (2012). [PubMed: 23162000]
94. Tobler PN, Fiorillo CD & Schultz W Adaptive coding of reward value by dopamine neurons. Science 307, 1642–1645 (2005). [PubMed: 15761155]
95. Behrens TEJ, Woolrich MW, Walton ME & Rushworth MFS Learning the value of information in an uncertain world. Nat. Neurosci 10, 1214–1221 (2007). [PubMed: 17676057]
96. 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, (2012).
97. 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). [PubMed: 29130693]
98. Myers JL & Sadler E Effects of range of payoffs as a variable in risk taking. J. Exp. Psychol 60, 306–309 (1960). [PubMed: 13727221]
99. Busemeyer JR & Townsend JT Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment. Psychol. Rev 100, 432–459 (1993). [PubMed: 8356185]
100. Gershman SJ Uncertainty and exploration. Decision (2019). doi:10.1101/265504
101. Frank MJ, Doll BB, Oas-Terpstra J & Moreno F Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. Nat. Neurosci 12, 1062–1068 (2009). [PubMed: 19620978]
102. Humphries MD, Khamassi M & Gurney K Dopaminergic Control of the Exploration-Exploitation Trade-Off via the Basal Ganglia. Front. Neurosci 6, 9 (2012). [PubMed: 22347155]
103. Pezzulo G, Rigoli F & Friston KJ Hierarchical Active Inference: A Theory of Motivated Control. Trends Cogn. Sci 22, 294–306 (2018). [PubMed: 29475638]
104. Botvinick M & Toussaint M Planning as inference. Trends Cogn. Sci 16, 485–488 (2012). [PubMed: 22940577]
105. FitzGerald THB, Dolan RJ & Friston K Dopamine, reward learning, and active inference. Front. Comput. Neurosci 9, 136 (2015). [PubMed: 26581305]
106. Friston KJ et al. Dopamine, affordance and active inference. PLoS Comput. Biol 8, e1002327 (2012). [PubMed: 22241972]
107. Weele CMV et al. Dopamine enhances signal-to-noise ratio in cortical-brainstem encoding of aversive stimuli. Nature 563, 397–401 (2018). [PubMed: 30405240]
108. 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). [PubMed: 18400958]
109. Gershman SJ, Norman KA & Niv Y Discovering latent causes in reinforcement learning. Current Opinion in Behavioral Sciences 5, 43–50 (2015).
110. Gershman SJ, Monfils M-H, Norman KA & Niv Y The computational nature of memory modification. Elife 6, (2017).
111. Redish AD, 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). [PubMed: 17638506]
112. Gardner MPH, Schoenbaum G & Gershman SJ Rethinking dopamine as generalized prediction error. Proc. Biol. Sci 285, (2018).
113. Gershman SJ The Successor Representation: Its Computational Logic and Neural Substrates. J. Neurosci 38, 7193–7200 (2018). [PubMed: 30006364]
114. Le Bouc R et al. Computational Dissection of Dopamine Motor and Motivational Functions in Humans. J. Neurosci 36, 6623–6633 (2016). [PubMed: 27335396]

115. Walton ME & Bouret S What Is the Relationship between Dopamine and Effort? Trends Neurosci. 42, 79–91 (2019). [PubMed: 30391016]

116. Westbrook A & Braver TS Dopamine Does Double Duty in Motivating Cognitive Effort. Neuron 91, 708 (2016).

117. Niv Y, Daw ND, Joel D & Dayan P Tonic dopamine: opportunity costs and the control of response vigor. Psychopharmacology 191, 507–520 (2007). [PubMed: 17031711]

Box 1: Temporal difference learning with belief states

Most reinforcement learning algorithms, including temporal difference (TD) learning,[15] 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.
- 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 modeled as a *partially observable Markov decision process,*[15,16] 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:

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

where P(s) is the prior over states.

The TD algorithm can be applied directly to the belief state representation using 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 RPE:

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

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 linear function approximation for clarity, nonlinear approximations are also possible.

Box 2: A unifying view of state representation

The "standard" TD model applied to dopamine 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 post-stimulus 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- and temporally-tuned neurons. Though simple and widely used, the CSC has been criticized for making incorrect predictions.[13,44,72,73]

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 into 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,[64] where the time bins correspond to "sub-states"; the key difference from the standard TD model is that the transition probabilities between sub-states 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 whose width increases and amplitude decreases as a function of time.[44,45,64] The discrete-time Markov approximation of the semi-Markov model offers one way of deriving these basis functions. If the uncertainty about the sub-state grows as a function of time, then the belief state will become increasingly spread out across multiple sub-states, 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.[47]

Box 3: Kalman temporal difference learning

Here we present a simplified version of the Kalman TD model.[47,75] The posterior over function approximation weights is Gaussian with mean $\widehat{w}$ and covariance matrix Σ. Similarly to the standard TD model (see Box 1), the Kalman TD model updates the mean using an RPE signal:

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

where $a = \Sigma \cdot h$ 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 RPE $\bar{\delta} = \delta / \lambda$ is normalized by the marginal variance $\lambda = h^{\mathsf{T}} \cdot a + \sigma^2$, where σ^2 is the variance of the reward distribution. This hypothesized normalization is consistent with data indicating that uncertainty rescales RPEs.[94] Greater unpredictability of rewards will increase λ and thus decrease the learning rate, while greater subjective uncertainty about the weights (encoded by a) will increase the learning rate, consistent with studies showing that high volatility of cue-reward associations leads to faster learning.[95]

The posterior covariance is updated according to:

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

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.

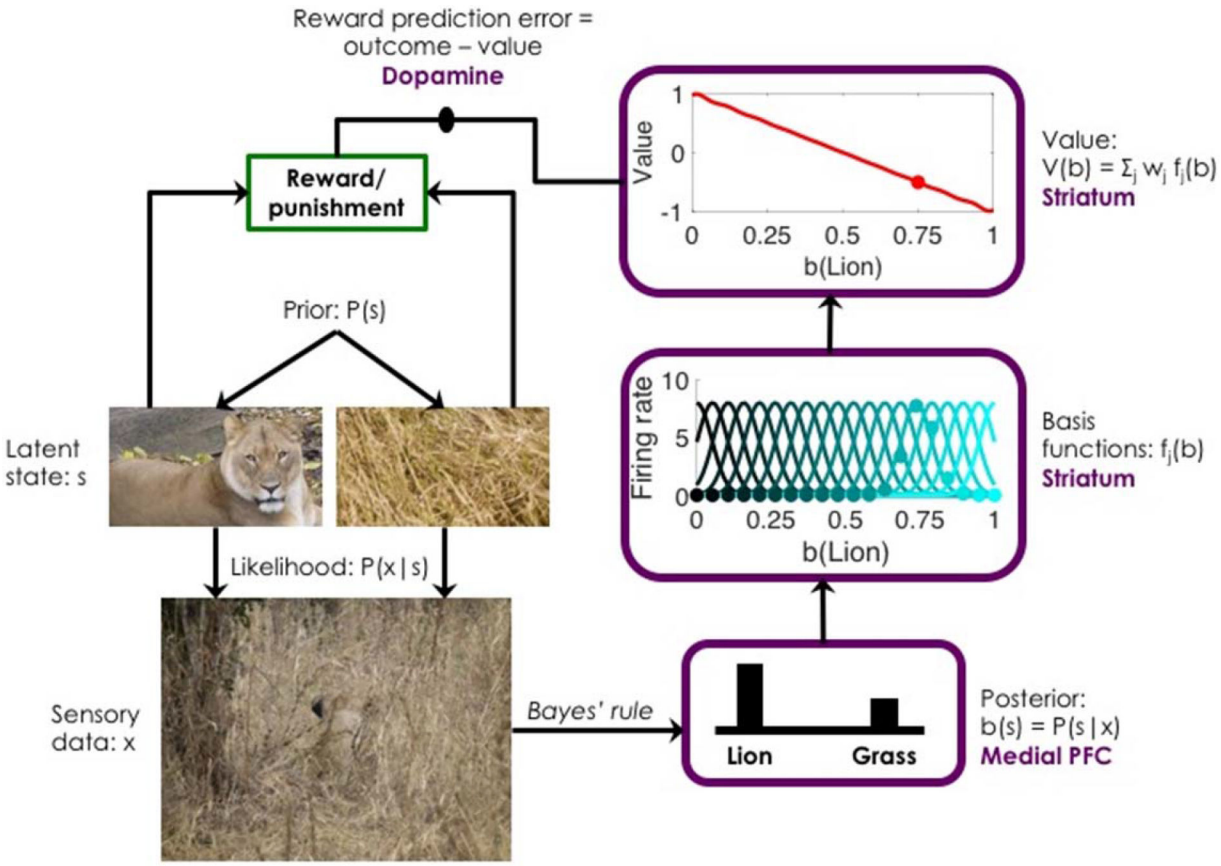

Figure 1. Schematic of the neural architecture for reinforcement learning under state uncertainty.

Bayesian inference combines noisy sensory data with a prior over latent states to compute the posterior distribution, or belief state, hypothesized to be encoded in the medial prefrontal cortex (PFC). The belief state is mapped into a distributed state representation (basis functions) in the striatum, which is in turn mapped onto a value function. Dopamine drives updating of the value function parameters by reporting a reward prediction error (the difference between observed and expected reward, or value).

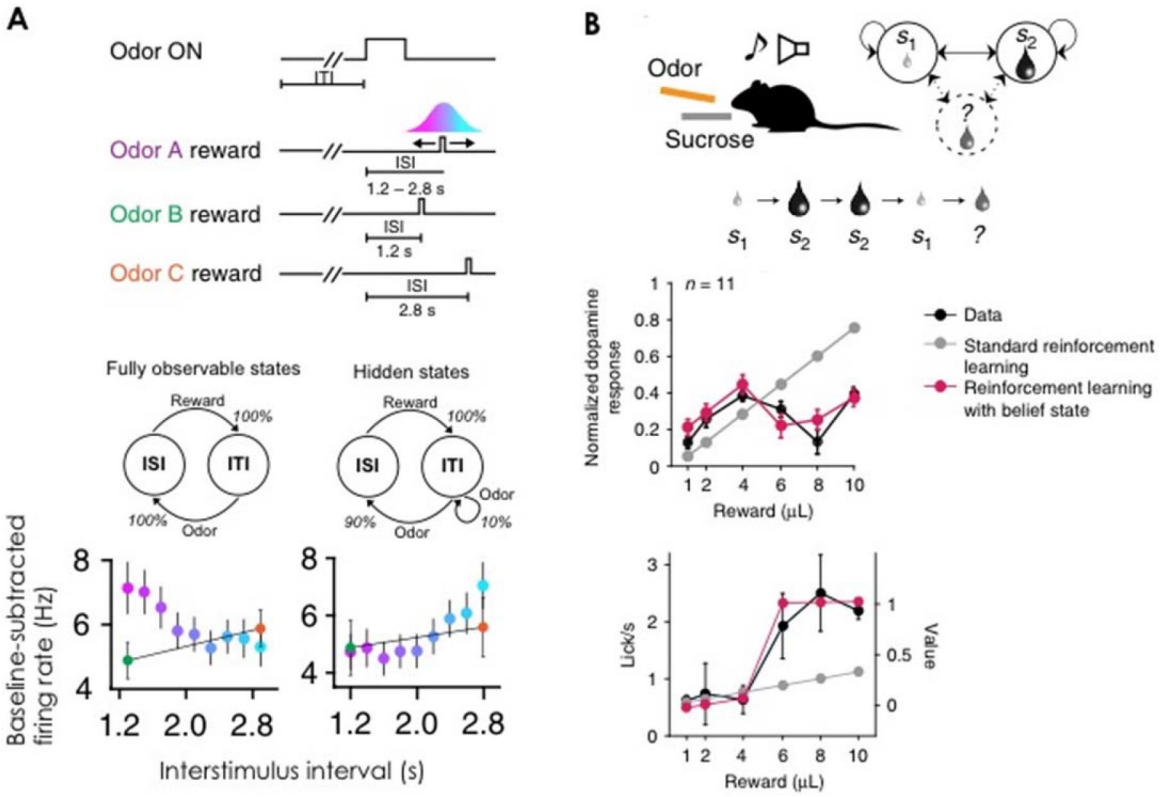

Figure 2. Experimental evidence for reflections of state uncertainty in dopamine signals.

(A) Experimental tasks and results from [64]. Mice observed an odor followed by a water reward. Odor A was associated with a variable odor-reward interval, whereas odors B and C were associated with fixed intervals. ISI: interstimulus interval between odor and reward. ITI: intertrial interval. The middle plots show the structure of the task as a probabilistic graphical mode. The bottom plots show the baseline-subtracted firing rates of optogenetically identified dopamine neurons in the ventral tegmental area. (B) Experimental task and results from [68]. Mice observed an odor followed by a water reward whose magnitude varied across blocks. The middle plot shows the normalized calcium response from dopamine neurons in the ventral tegmental area measured using fiber photometry. The bottom plot shows anticipatory licking and the predicted values. Animals were trained using blocks of either small or big reward trials first. In rare trials (probe trials), animals received intermediate-size reward. The x-axis indicates the magnitudes of reward in the probe trials.

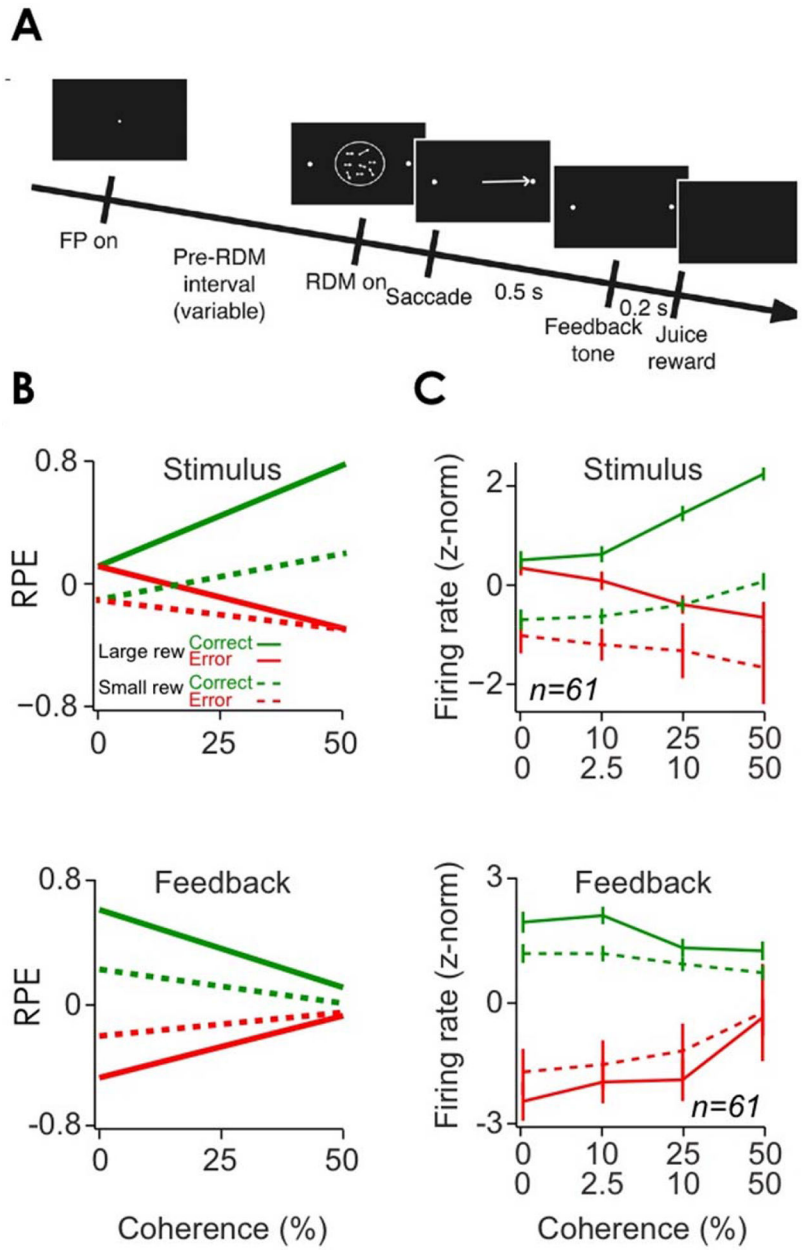

Figure 3. Experimental evidence for uncertainty-dependent dopamine signals in a perceptual decision making task.

(A) Experimental task from [68,69]. Monkeys observe randomly moving dots and then make a judgment about their direction. The proportion of coherently moving dots is manipulated across trials. (B) Predictions from the belief state model. At stimulus onset, reward prediction error (RPE) response increases as a function of coherence on correct trials, but decreases as a function of coherence on error trials. This pattern is inverted at feedback onset. (C) Recordings of dopamine neurons under the same condition as in (B), confirming the theoretical predictions.

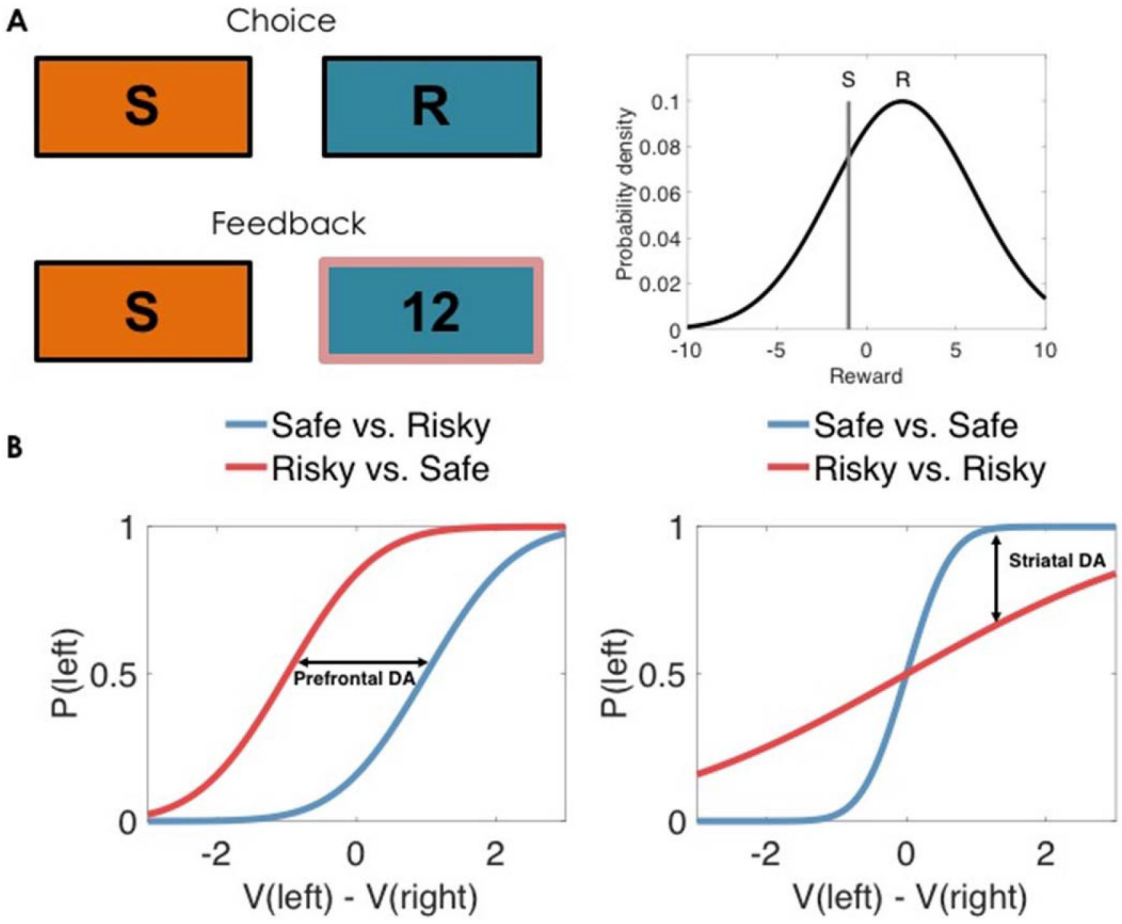

Figure 4. Two forms of uncertainty have distinct effects on exploratory choice, and are governed by distinct dopamine afferents.

(A) A two-armed bandit task in which each arm is either "safe" (deterministic) or "risky" (stochastic). (B) Schematic of 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 (DA) levels. Evidence suggests that the magnitude of the uncertainty bonus is controlled by prefrontal dopamine (DA) levels. The right plot illustrates the manipulation of total uncertainty: when the 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 DA levels.