nature neuroscience

Perspective

https://doi.org/10.1038/s41593-024-01705-4

Explaining dopamine through prediction errors and beyond

Received: 24 April 2023

Accepted: 13 June 2024

Published online: 25 July 2024

Check for updates

Samuel J. Gershman[1,2] ✉, John A. Assad[3], Sandeep Robert Datta[3], Scott W. Linderman[4], Bernardo L. Sabatini[2,3,5], Naoshige Uchida[6] & Linda Wilbrecht[7]

The most influential account of phasic dopamine holds that it reports reward prediction errors (RPEs). The RPE-based interpretation of dopamine signaling is, in its original form, probably too simple and fails to explain all the properties of phasic dopamine observed in behaving animals. This Perspective helps to resolve some of the conflicting interpretations of dopamine that currently exist in the literature. We focus on the following three empirical challenges to the RPE theory of dopamine: why does dopamine (1) ramp up as animals approach rewards, (2) respond to sensory and motor features and (3) influence action selection? We argue that the prediction error concept, once it has been suitably modified and generalized based on an analysis of each computational problem, answers each challenge. Nonetheless, there are a number of additional empirical findings that appear to demand fundamentally different theoretical explanations beyond encoding RPE. Therefore, looking forward, we discuss the prospects for a unifying theory that respects the diversity of dopamine signaling and function as well as the complex circuitry that both underlies and responds to dopaminergic transmission.

An important turning point in the study of dopamine was the application of engineering concepts from reinforcement learning (RL) theory (Box 1). This theory formalized the computational problem that needed to be solved (predicting future reward) and an algorithm (temporal difference learning) for solving it. The biological plausibility of this algorithm was first indicated by data showing that the phasic firing of dopamine neurons resembled the reward prediction error (RPE) signal used by the learning algorithm to update reward predictions[1,2].

Subsequent work tested the RPE hypothesis of dopamine quantitatively[3–5], validated it causally by showing that perturbations of dopamine affect learning[6,7] and identified cellular mechanisms by which the error computation and downstream plasticity modulation might be implemented[8,9]. These lines of work have been thoroughly reviewed elsewhere[10,11]. Our focus here is on empirical data that appear to challenge the RPE hypothesis and to what extent does it stand up to these challenges.

Answering this question is tricky because the RPE hypothesis is not really one model but rather a family of models that share some key ideas. RL theory allows us to define the underlying assumptions about the environment and model architecture in different ways while still adhering to the framework of temporal difference learning. We could try to search through the space of viable models to identify those that match all of the empirical data. However, doing so would forfeit some of the explanatory thrust that comes with a normative framework—by understanding why nature may have selected one architecture over another, we can reveal deeper principles about brain function.

[1]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA, USA. [2]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA, USA. [3]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [4]Department of Statistics and Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA, USA. [5]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [6]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA. [7]Department of Psychology and Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA. ✉e-mail: gershman@fas.harvard.edu

https://doi.org/10.1038/s41593-024-01705-4

BOX 1

Brief review of RL concepts

The goal of an RL agent is commonly taken to be the maximization of cumulative future reward, or value:

$$V_t = E_\pi(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots),$$

where $\pi(a|s)$ is a policy mapping the agent's state s to a probability distribution over action a, r_t is the reward received at time t, γ is a temporal discount factor governing how far into the future the agent cares about accumulating rewards and $E_\pi(\cdot)$ is an expectation with respect to the policy as well as any randomness in state transitions and rewards. Note that V_t is a short-hand for $V(s_t)$; that is, value is a state-dependent variable (as is r_t).

When state transitions and rewards only depend on the current state and action, the environment is referred to as a Markov decision process. In this case, the value function can be expressed recursively (the Bellman equation) as

$$V_t = E_\pi\left(r_t + \gamma V_{t+1}\right).$$

An agent who has learned an estimate of the value, $\hat{V}_t$, can use the Bellman equation to measure the inconsistency in values between consecutive time points as

$$\delta_t = r_t + \gamma \hat{V}_{t+1} - \hat{V}_t.$$

This is the temporal difference error, which constitutes the formal version of the RPE hypothesis, namely that phasic dopamine reports δ_t.

The temporal difference error specifies how the value estimate should be updated based on reward feedback. Assume that the value function is governed by a set of modifiable parameters (w). These parameters should be updated to reduce error:

$$\Delta w = \alpha \delta_t \nabla_w \hat{V}_t,$$

where α is a learning rate and $\nabla_w \hat{V}_t$ is the gradient of the value estimate with respect to the parameters. A common assumption is that the value estimate is linear.

$$\hat{V}_t = \sum_d w_d f_d(s_t),$$

where $f_d(s_t)$ is a feature function defined over states. In this case, the gradient is equal to the vector of features—$\nabla_w \hat{V}_t = f(s_t)$. Intuitively, the parameters correspond to feature weights, which are increased when errors are positive and decreased when errors are negative.

Answering the 'why' question requires us to analyze the computational rationale of RL algorithms and their neural implementation.

In the rest of this Perspective, we consider several recent empirical challenges to the RPE hypothesis, which have led to alternative hypotheses about the function of dopamine. We argue that each challenge can be met by suitably modifying the RPE hypothesis. Notably, each of these modifications is motivated by a computational rationale rather than ad hoc data fitting. We also highlight how these modifications have led to new predictions that have been tested experimentally. Finally, we consider several empirical challenges that have not yet been met by the RPE hypothesis and how new models attempt to address these challenges with the aid of new measurement technologies and analysis methods.

Why does dopamine ramp up as animals approach reward?

In a seminal experiment, details in ref. 12 showed that extracellular dopamine concentration in the striatum slowly ramps up as rats approach the reward location in T-maze. This was initially interpreted as problematic for the RPE hypothesis because it was not clear why the RPE should exhibit a ramp[12–14]. By contrast, it was easy to see how a ramp would be consistent with a value signal, which necessarily increases as a reward is approached. This value interpretation was reinforced by subsequent studies suggesting that dopamine ramps covary with reward rate and response latency, a behavioral proxy for motivation[15–17]. Dopamine ramps also anticipate (and causally control) the timing of reward-related movements[18,19], suggesting that they may have a role in action evaluation rather than prediction error signaling.

Dopamine ramps may or may not be consistent with the RPE hypothesis, depending on the shape of the value function. To understand this, note that when the animal is approaching reward (hence $r_t = 0$) and the discount factor is close to 1, the RPE is approximately the time derivative of value—$\delta_t \approx \hat{V}_{t+1} - \hat{V}_t$. This means that the RPE will ramp whenever the value function approximation is convex (its slope is increasing with proximity to reward)[20]. More generally, the RPE will ramp for discount factors below 1 when the value function approximation is 'convex enough' to compensate for temporal discounting. One contributor to convexity is the function approximator; for example, if the features being used to approximate the value function are more temporally imprecise farther away from the reward, thereby creating a distance-dependent 'blur' of the value function. Another contributor to convexity is state uncertainty, as we describe below.

The time derivative view shows that, in principle, RPEs can ramp, but it does not show that this is in fact the underlying explanation of dopamine ramps. In a series of studies, Kim et al. used virtual reality to carefully separate the predictions of value and RPE accounts, showing repeatedly that the dynamics of dopamine were more consistent with the RPE account[21]. For example, virtually teleporting a mouse over varying distances to the same location on a track should result in the same value signal but different RPEs, because longer distances should generate larger RPEs, as seen in the experimental data (Fig. 1). These findings generalized across measurements of calcium in dopamine neuron cell bodies and axons projecting to the ventral striatum, as well as measurements of spiking activity in optogenetically identified dopamine neurons. Finally, quantitative model comparison strongly favored the RPE account.

Despite this empirical support for the RPE account, the studies discussed in ref. 21 did not address a lingering puzzle—why do ramps mainly appear in spatial navigation tasks and not in Pavlovian conditioning tasks? A key difference between these tasks is the continual acquisition of sensory feedback during spatial navigation but not during the interstimulus interval of a Pavlovian conditioning task. Because timekeeping in the brain is noisy, an animal's subjective estimate of objective time will become increasingly uncertain as time elapses in the absence of sensory feedback. If τ is the noisy internal timekeeping signal and $p(t|\tau)$ is the posterior probability distribution over objective time given the timekeeping signal (the 'uncertainty kernel', that is, the distribution of actual times that are likely given the animal's estimation), then the optimal value estimate is an average of value estimates at each hypothetical objective time weighted by its probability:

$$\hat{V}_\tau = \sum_t p(t|\tau)\hat{V}_t = E(\hat{V}_t|\tau).$$

In a Pavlovian task, the true value function, V_t, is a convex and increasing function of time because as time elapses, the reward draws nearer. If the approximation $\hat{V}_t$ is convex as well, Jensen's inequality implies that $\hat{V}_\tau \geq \hat{V}_t$. This means that $\hat{V}_\tau$ will tend to overestimate the

https://doi.org/10.1038/s41593-024-01705-4

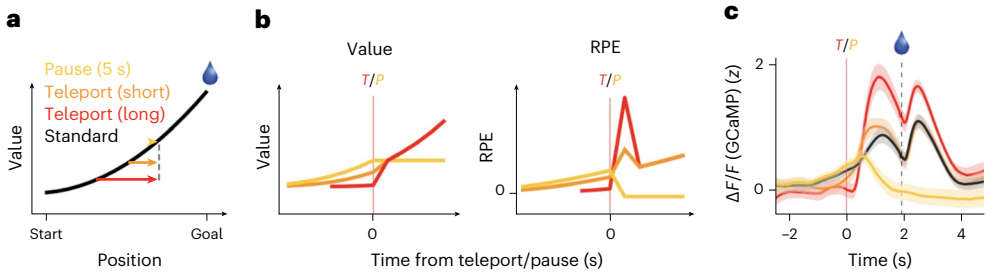


Fig. 1 | Dopamine ramps reflect RPE, not value. a, Experimental design used in ref. 21. Mice in a virtual linear track were exposed to the following four different conditions: pausing for 5 s at a particular position on the track, teleporting to that position from a short distance away, teleporting from a long distance away or arriving at the position following a 'standard' (constant speed) sequence of positions. **b**, If dopamine reports value, then its response at the probed position should be the same across conditions. By contrast, if dopamine reports RPE, then its response should be greatest after the long teleport, which produces the largest change in value. **c**, The RPE hypothesis was supported by fiber photometry measurements of calcium signals in dopamine axons projecting to the ventral striatum. Reproduced with permission from ref. 21,Cell Press.

true value V_t. However, the true value can still be learned asymptotically by temporal difference learning (with a sufficiently powerful function approximator), provided the uncertainty kernel is known and the value estimate $\hat{V}_T$ does not change abruptly over short intervals. This stability condition is satisfied by Pavlovian conditioning tasks but is violated by spatial navigation tasks due to the sensory feedback that updates the uncertainty kernel and thus the value estimate. As a consequence, temporal difference learning will produce biased value estimates even as the RPEs asymptotically go to 0.

The analysis in ref. 22 showed that the bias could be corrected by modifying the error-driven learning rule to decay value estimates in proportion to the informativeness of sensory feedback. This correction drives the bias to zero but can produce asymptotically nonzero RPEs (Fig. 2). In particular, spatial navigation tasks typically produce a ramping profile of RPEs. The value decay model provides a normative foundation for similar mechanisms previously proposed to account for ramping[23], as well as behavioral data from humans and rodents[24–26]. In Pavlovian tasks, there is no sensory feedback, and therefore the decay term is 0 (that is, no ramping), explaining why ramps do not typically occur in Pavlovian tasks. In these tasks, there is no bias that needs to be corrected.

An important feature of bias correction is that the RPEs converge to a steady state that is proportional to value, where the constant of proportionality depends on the informativeness of sensory feedback. This sheds light on conflicting interpretations of dopamine in the literature. The value interpretation of dopamine ramps may in fact be consistent (at least asymptotically) with the RPE account. The value decay model adds nuance to this unification by restricting the conditions under which RPEs resemble value signals—when sensory feedback is relatively uninformative (for example, in Pavlovian tasks), RPEs no longer look like value signals, instead converging to 0.

According to the value decay model, the shape of ramps depends on the temporal profile of sensory feedback—ramps can be increasing, decreasing, flat or even nonmonotonic. A study reported in ref. 22 manipulated the temporal profile by gradually darkening the visual scene. Consistent with model predictions, this produced a dopamine bump rather than a ramp. The key idea was that darkening initially increases the convexity of the value function approximation and then later causes it to become concave (due to the blurring of prereward and postreward values). This result illustrates how understanding the computational rationale of ramping can yield new and nontrivial predictions.

Alternative models for dopamine ramps have also been proposed[27,28]. Hamid et al. proposed an architecture in which different dorsal striatum subregions implement 'experts' that predict future reward[28]. Dopamine transmission in each subregion encodes the accumulated evidence for the accuracy of the corresponding expert's predictions, thereby mediating credit assignment. The dorsomedial striatum was posited as an action-outcome predictor based on the distance to reward, and accordingly, dopamine in this region was interpreted as an evidence accumulator for the controllability of reward. This model explained why dopamine in the dorsomedial striatum ramped up in an instrumental task (high controllability) but ramped down in a Pavlovian task (low controllability).

By contrast, dopamine signals originating in the ventral tegmental area and their projections to the ventral striatum do not depend on action-outcome controllability as long as sensory cues indicate the distance to a goal[21,29]. In the absence of such cues, the same dopamine signals depend on controllability[29]. These results suggest that mesolimbic dopamine signals use multiple sources of information to determine progress toward a goal, which could be compatible with several theoretical accounts. A preprint from ref. 29 argues that the rapid appearance of ramps probably reflects the use of an internal model to compute goal progress rather than temporal difference learning, which requires multiple trajectories to propagate RPEs from the goal toward earlier predictive cues[30].

Why does dopamine respond to sensory features?

It has long been known that dopamine signaling increases in response to new or surprising stimuli, independent of their motivational value[31–34]. There have been a number of theoretical efforts to reconcile these observations with the RPE hypothesis. Kakade and Dayan argued that novelty responses might reflect an exploration bonus added to the RPE, possibly due to the optimistic initialization of value estimates[35]. In the setting of threat learning, value estimates may also be initialized pessimistically to discourage exploration of potentially dangerous states. Evidence for this kind of novelty penalty has been shown in studies of dopamine projections to the tail of the striatum[36,37]. An elevated response to new stimuli has also been reported in the ventral tegmental area (which projects to the ventral striatum) and the substantia nigra pars compacta (which projects to the dorsal striatum)[38]. One caveat for interpreting these studies is that some apparent novelty responses could reflect generalization across stimuli in the environment (particularly in rich environments where stimuli tend to be paired with reward)[39,40], rather than a novelty bonus or penalty per se. Further work is needed to disentangle these possibilities.

Another possibility is that dopamine responses reflect an intrinsic preference for information signaled by stimuli, such that RPEs are elevated when the animal receives more information than expected[41,42]. This may explain the recent finding that some dopamine neurons do not respond to new stimuli after exposure to many other new stimuli[43].

None of these models make sense of more recent data indicating a role for phasic dopamine signals in learning sensory predictions. For example, rats can learn an association between two neutral stimuli

https://doi.org/10.1038/s41593-024-01705-4

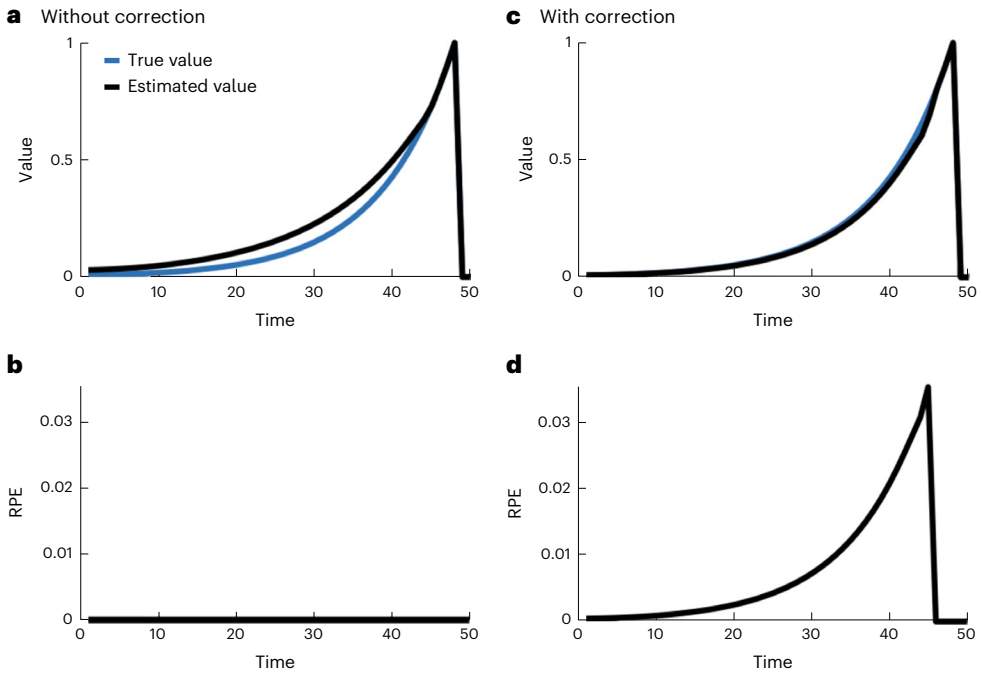


Fig. 2 | The role of state uncertainty in ramping. a, A standard temporal difference learning algorithm will systematically overestimate value under conditions where state uncertainty is dynamically resolved by sensory feedback. **b**, Despite systematic overestimation of value, the RPE can still converge to 0. **c**, Value overestimation can be corrected by decaying values in proportion to the informativeness of sensory feedback. **d**, The correction induces a ramping RPE. Reproduced with permission from ref. 22, Cell Press.

(A and B), which later allows them to generalize conditioned responses to A after pairing a reward with B, an effect known as sensory preconditioning. Inhibiting dopamine neurons during the A–B preconditioning phase eliminates this effect, while activating them strengthens it[44]. Rats are also sensitive to changes in the identity of reinforcers (for example, the flavor of food pellets) during learning. If A is paired with reward and then subsequently paired with reward in compound with another stimulus B, the new stimulus fails to elicit conditioned responding (the classic Kamin blocking effect), but this blocking effect is attenuated if the reinforcer is changed to one with equal incentive value, a phenomenon known as identity unblocking. Inhibiting dopamine neurons during the compound reinforcement phase eliminates identity unblocking, again supporting the view that phasic dopamine has a causal role in learning sensory predictions[45]. Electrophysiological recordings paint a consistent picture—dopamine neuron spiking is elicited by surprising changes in sensory features of reinforcers[46].

One way to understand these findings is to posit that dopamine is a learning mechanism for a form of model-based RL[47]. On this account, dopamine signals update an internal state transition and reward model, which then guides behavior through mental simulation and planning. Alternatively, dopamine responses have been conceptualized as a form of 'generalized' prediction error. According to this hypothesis, dopamine performs the same error computation for reward and sensory features[48]. The computational rationale for this hypothesis comes from the successor representation model (Box 2), which offers an alternative algorithm for estimating values[49,50]. The essential idea is that the brain learns a 'predictive map' of expectations about the frequency of sensory features in the near future (possibly represented in the hippocampus[51]). The predictive map is analogous to the value function (a discounted sum of future events) but applied to arbitrary sensory features.

The generalized prediction error hypothesis represents a substantial departure from the RPE hypothesis in that the error signal is vector-valued rather than scalar. This opens the door to partially explaining the diverse tuning profiles of dopamine neurons[52–55]. It also raises the question of how the error vector is encoded in the population. One possibility is a kind of labeled-line code where anatomically segregated pathways convey different parts of the error vector to distinct targets. A number of studies support this hypothesis, showing that striatal dorsal/ventral, medial/lateral and rostral/caudal targets receive different information from dopamine signals[36,37,56–61] (Box 3). Increases in dopamine in some of these areas are time-locked to both positive and negative or aversive events and may encode prediction errors with different purposes. For example, dopamine transmitted to the tail of the striatum has been proposed to carry a 'threat prediction error' that is suppressed when rodents explore a nonthreatening new object[37]. Stronger novelty responses in this dopamine projection were inversely correlated with object exploration, and ablating the projection promoted object exploration. This example is particularly interesting because it suggests a way in which the generalized prediction error hypothesis might be convergent with the exploration bonus hypothesis proposed in ref. 35.

Although there is some evidence for large-scale anatomical segregation of different error signals, there do not seem to be enough labeled lines to support a sufficiently rich generalized prediction error. Another possibility is that the error vector is mixed at the population level, necessitating a downstream 'demixing' process. Supporting this hypothesis is evidence that reinforcer identity can be decoded from ensembles of dopamine neurons but not from single neurons[62] (Fig. 3). Notably, identity is decodable only early during a block of trials following an identity change. This is consistent with the hypothesis that identity information is conveyed by an error signal that diminishes over the course of learning. The apparently distributed nature of identity information raises important questions about how downstream systems decode and use this information. It also raises the question of why reward magnitude can be decoded from individual neurons but identity cannot. We conjecture that reward magnitude should be privileged in any distributed representation because it is particularly relevant for value estimation; single neuron decodability may reflect the robust encoding of magnitude information.

https://doi.org/10.1038/s41593-024-01705-4

BOX 2

The successor representation

The value function can be decomposed into an inner product between a predictive map (the successor representation, M) and the reward function (R):

$$V_t = \sum_s M_t(s)R(s),$$

where s denotes the state, t denotes time and M is defined as the expected discounted state occupancy:

$$M_t(s) = E_\pi(I_t(s) + \gamma I_{t+1}(s) + \gamma^2 I_{t+2}(s) + \ldots)$$

and $I_t(s) = 1$ if $s_t = s$ (0 otherwise). Intuitively, the successor representation encodes how often each state will be visited in the near future, where the effective time horizon for prediction is determined by the discount factor γ.

Like the value function, the successor representation obeys a Bellman equation in Markovian environments:

$$M_t(s) = E_\pi\left(I_t(s) + \gamma M_{t+1}(s)\right)$$

Consequently, an estimate $\hat{M}_t(s)$ can be updated using temporal difference learning:

$$\Delta\hat{M}_t(s) = \alpha\delta_t(s),$$

where α is again a learning rate, and the prediction error is vector-valued (that is, one value for each possible state):

$$\delta_t(s) = I_t(s) + \gamma\hat{M}_{t+1}(s) - \hat{M}_t(s).$$

This 'tabular' approximation can be straightforwardly extended to the function approximation setting (as in Box 1). In essence, the update is the same computation as described in Box 1 for scalar RPEs, but now applied to state occupancy.

BOX 3

Origins of functional diversity

Dopamine has a diversity of effects on different parts of the brain[101]. This diversity could arise from differences in signal content, differences in downstream computation or both. Here we discuss recent insights into the functional specificity of different dopamine projections.

Hughes et al. analyzed the activity of dopamine neurons in the ventral tegmental area while head-fixed mice were given a reward every 10 s[102]. The researchers identified three populations of dopamine neurons that encoded distinct components of the 'impulse vector' (forces exerted by the animals). One population increased firing for backward movement and decreased firing for forward movement at the time of reward delivery; a second population exhibited the opposite pattern; a third population was similar to the second population but more spread out over time, with little phasic modulation at the time of reward.

Using a cued reward-seeking task, a study in ref. 103 identified separate populations of dopamine neurons in the ventral tegmental area with topographically organized selectivity—neurons in the lateral division (projecting to the lateral shell of the ventral striatum) resembled classical RPE neurons, whereas neurons in the medial division (projecting to the medial shell of the ventral striatum) exhibited sustained activity between cue presentation and reward delivery, suggesting a representation of motivational state.

These examples illustrate cases in which functional specificity arises from differences in signal content. The same signals may also mediate different consequences depending on their targets. Notably, striatal and prefrontal neurons expressing D1 receptors (D1Rs) versus D2 receptors (D2Rs) respond differently to dopamine inputs. In the striatum, neurons expressing D1Rs promote action via the direct pathway, whereas neurons expressing D2Rs inhibit action via the indirect pathway[104]. In the prefrontal cortex, D1R-expressing neurons promote memory maintenance, whereas D2R-expressing neurons promote memory updating[105].

Recent data have highlighted ways in which cell-type specificity and projection specificity interact to produce functional diversity. For example, a septal-hypothalamic pathway mediating aggression is modulated by a population of dopamine neurons projecting from the ventral tegmental area to D2R-expressing neurons in the lateral septum[106]. Optogenetic activation of these dopamine neurons can evoke aggressive behavior, but this effect is prevented by D2R antagonism in the lateral septum. Another example is the encoding of information about reward distributions through the opponency of neurons expressing D1Rs versus D2Rs in the striatum, which may interact with the functional differences of projections from distinct populations of dopamine neurons[107,108].

Lee et al. proposed a different way of understanding dopamine heterogeneity, which retains the vector idea but not the notion of generalized prediction error—the scalar RPE can be disassembled into multiple feature-specific channels and then additively reassembled by its targets[63]. This idea is motivated by the topography of projections from midbrain to striatal and cortical targets, as noted earlier in our discussion of labeled-line codes. Strictly speaking, only reward-predictive features should elicit dopamine responses according to this account; however, in ref. 63, it has been shown that when the function approximator is a deep neural network, reward-irrelevant feature selectivity 'leaks' into dopamine responses due to the distributed nature of the state representation. More work is required to determine whether this kind of model can explain the full breadth of findings related to sensory feature coding by dopamine neurons.

Why does dopamine respond to motor features?

In addition to encoding sensory features, dopamine neurons encode motor-related variables such as initiation, vigor and kinematics[52,64–66]. At least some aspects of this motor encoding are not well-explained by the RPE hypothesis[67]. A possible account might be the one already given above for understanding sensory encoding—dopamine neurons signal generalized prediction errors for motor features. This would make sense if we assume that the learned predictive representation is used for action selection, in which case it would need some mechanism for encoding motor features. However, some studies find that activity in dopamine projections to the dorsal striatum does not reliably distinguish between movement types[64,68], suggesting a fairly coarse feature encoding. This is problematic if one assumes that the encoded features are being used downstream for action selection. To some extent, this coarse representation may be refined by interaction with specific glutamate signals impinging on striatal neurons and/or local filtering of signals at the level of individual glutamate and dopamine axon terminals[69,70]. Synaptic-level resolution in vivo imaging methods will be needed to resolve this issue.

https://doi.org/10.1038/s41593-024-01705-4

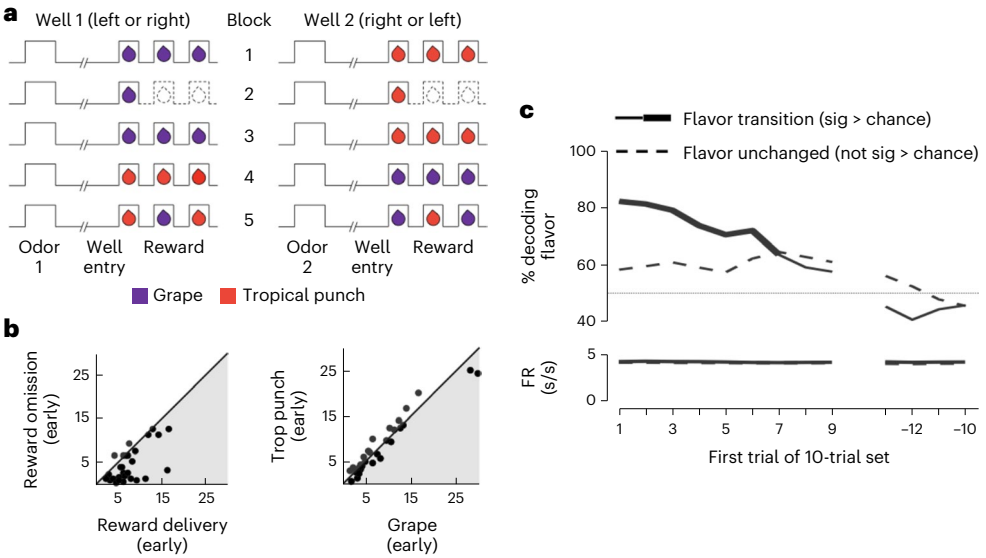


Fig. 3 | Dopamine ensembles encode generalized prediction errors. **a**, Experimental design used in ref. 62. On some block transitions, reward magnitude changed (omission trials are indicated by dashed circles). On other block transitions, reward identity (flavor) changed. **b**, Early during a block, putative dopamine neuron FRs, recorded from the ventral tegmental area, discriminated between reward magnitude (left) but not reward identity (right). Each dot corresponds to a single neuron. **c**, A classifier applied to the ensemble of neurons (top) could discriminate between reward identities early (but not late) after an identity change, consistent with a distributed encoding of generalized prediction error that diminishes over the course of a block. The bottom plot shows the average baseline FR. Adapted from ref. 62, CC0 (http://creativecommons.org/publicdomain/zero/1.0/). FR, firing rate.

Another complicating factor is that motor control in the brain is fundamentally distributed—multiple regions apart from the basal ganglia, such as the motor cortex and cerebellum, exert semi-independent control over motor behavior. This observation poses a challenge for learning algorithms that update parameters in a particular region based on the aggregate policy—the RPE used to update parameters is shared across all of these regions, resulting in a credit assignment problem. If the basal ganglia promote action A but action B is selected due to the influence of other regions, the basal ganglia parameters may be incorrectly assigned credit or blame.

'Off-policy' algorithms such as Q-learning can finesse this problem by learning about a policy that is different from the one generating behavior. When suitably parametrized, Q-learning can be applied to the continuous action spaces typical in motor control problems[71]. Critically, the teaching signal for updating depends on both the classical RPE and an 'action surprise' signal (the deviation of the sampled action from the highest-probability action). Because this action surprise signal is nonspecific (it can be evoked by any low-probability action), it can explain the finding that motor-evoked dopamine signals do not distinguish between movement types. It can also explain why dopamine responses to lever-pressing decrease with practice[64], as actions become more reliable and hence less surprising. Recent data have provided direct evidence for action prediction errors in dopamine projections to the tail of the striatum[72]. These signals were evoked when animals took familiar actions in response to an unfamiliar cue and were not modulated by reward value. Optogenetic stimulation of these projections at the time of choice induced a contralateral choice bias (consonant with earlier pharmacological studies) without modulating other movement parameters. Notably, stimulation at the time of outcome did not induce a choice bias, indicating that the signals were not conveying a traditional RPE.

What is the role of dopamine in action selection and motivation?

The RPE hypothesis posits that dopamine functions as a learning signal. Although this signal can influence action selection via updating of action values or policy parameters following reinforcement, the original hypothesis does not provide any role for dopamine in the action selection process itself. Nonetheless, studies have indicated such a role.

Acute optogenetic manipulations of dopamine at the time of action selection can alter an animal's behavior, including movement choice[68,73,74], vigor[15], initiation[18] and kinematics[75]. It should be noted, however, that some data argue against a role for phasic dopamine in real-time control of action selection. In particular, one study found that the effects of optogenetic manipulations were temporally restricted to the postreward period[76]. Another study found that optogenetic activation of dopamine axons caused a reduction in motivation[77].

Some RL models treat the learning and action selection functions of dopamine as separate channels, possibly implemented by phasic and tonic activity, respectively[78,79]. These channels must be linked—tonic activity is a time integral of phasic activity. Mathematically, it is possible to reconcile this constraint with the dual-channel hypothesis, under the assumption that tonic dopamine reports an estimate of the average reward and phasic dopamine reports an RPE that is referenced to the average reward[80] (Box 4). Critically, the estimate of average reward is simply an accumulator of the RPE. This is closely related to the average reward model of dopamine proposed in ref. 81, but has the advantage that the two channels are mutually consistent.

Average reward has a special role in action selection by calibrating the optimal level of effort. For example, an animal in an operant conditioning task with a fixed ratio schedule can earn more rewards by responding more frequently, but this incurs a physical effort cost. A higher average reward (for example, due to a higher ratio) means that the animal can earn more rewards for the same amount of effort and therefore should respond more frequently. This is indeed what happens with normal rats, but not with dopamine-depleted rats[82], a finding that is well-explained by an average reward model of tonic dopamine[83] and is consistent with other motivational deficits reported in dopamine-depleted animals[84]. In light of the average reward model, one possibility is that the motivational effects of optogenetic manipulations may be mediated by transient changes in tonic dopamine levels. Indeed, recent data suggest that interactions between tonic and phasic channels may be complex[85].

Some models have directly incorporated phasic dopamine into the action selection process[86,87]. For example, the influential opponent

https://doi.org/10.1038/s41593-024-01705-4

BOX 4

Linking phasic and tonic dopamine through average reward RL

The dual-channel hypothesis posits that phasic dopamine encodes RPEs and tonic dopamine encodes average reward. This can be formalized as follows, using the definition of differential value instead of discounted value:

$$V_t = E_\pi\left(r_t - \rho + V_{t+1}\right)$$

$$\rho = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} E_\pi(r_t),$$

where ρ is the asymptotic average reward over time. Wan et al.[80] proved that differential values can be estimated using a temporal difference learning algorithm:

$$\Delta\hat{V}_t(s) = \alpha\delta_t$$

$$\Delta\hat{\rho}_t = \eta\alpha\delta_t,$$

where $\hat{\rho}_t$ is an estimate of the average reward, $\eta > 0$ is a step-size parameter, α is a learning rate and δ_t is the average reward RPE:

$$\delta_t = r_t - \hat{\rho}_t + \hat{V}_{t+1}(s) - \hat{V}_t(s)$$

Note that, although individual RPEs may decrease toward 0 over the course of learning, the same is not necessarily true of the accumulated RPE (the estimate of average reward)—if there are initial positive RPEs, the accumulator will increase and then asymptote as RPEs go to 0. Critically, the same RPE is used to update both channels, making the model consistent with the physiological constraint that the same dopamine signal determines tonic and phasic activity at different timescales.

The average reward model formalizes the longstanding idea that tonic dopamine activity can exert an antagonistic effect on phasic activity through the action of autoreceptors[109]. The model predicts that as the reward rate increases, the amplitude of phasic responses to unexpected rewards should become progressively smaller[81], as observed experimentally[110].

actor learning model[86,88] assumes that phasic dopamine levels influence action selection by binding to striatal D1 (Go) receptors on direct pathway neurons and D2 (NoGo) receptors on indirect pathway neurons, with increases and decreases (or pauses) in dopamine and different binding affinities determining asymmetries in Go versus NoGo behavior. These models raise the question of why, from a normative perspective, the same dopamine signal should participate in both learning and action selection.

One attempt to resolve this question posits that dopamine has a role in goal-directed action planning[89]. RPEs, on this account, provide a directive signal for policy optimization, unifying planning, habit formation and learning. The key idea is that the same RPE signal can be reduced by improving value estimates (learning) in ventral striatum and improving the policy (planning and habit formation) in dorsal striatum. Positive RPEs delivered to dorsal striatum are hypothesized to drive increases in action intensity (vigor) via dopamine projections to striatal neurons expressing D1Rs. When action intensity is too high, resulting in forfeited reward, the ensuing negative RPE drives a decrease in action intensity (which could potentially happen at fast timescales, in addition to slower trial-by-trial adjustments). This account explains why vigor is modulated by acute manipulations of dopamine at the time of action—positive RPEs indicate that more reward can be earned by adopting higher action intensity[15].

Beyond RPEs

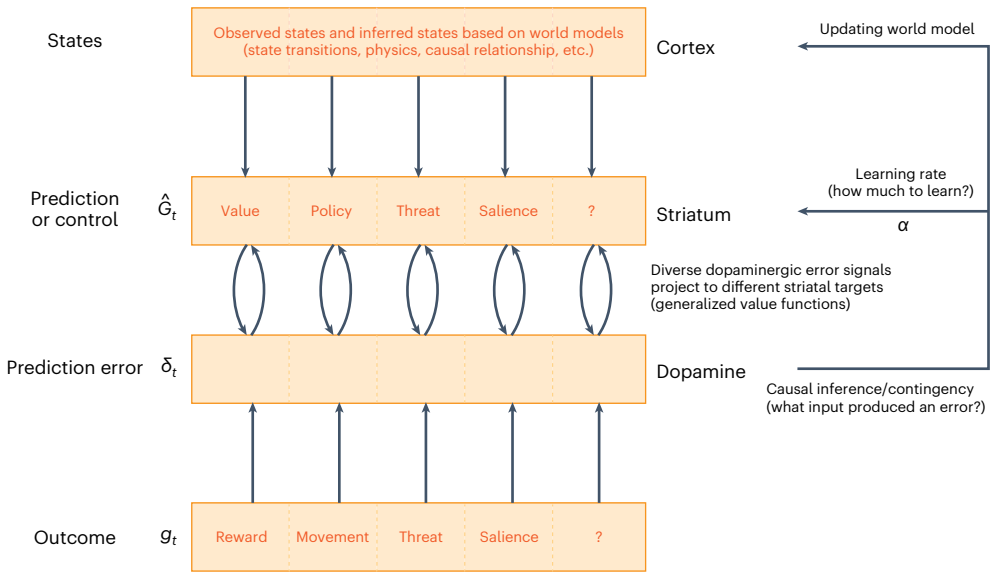
We have focused on explanations of dopamine function within an RL framework, showing that a suitably generalized concept of prediction error can explain a wide range of findings about dopamine. However, recent studies have presented data that might fundamentally challenge the prediction error hypothesis and suggest alternative theories.

Building on a sophisticated model of Pavlovian conditioning[90], studies in refs. 91,92 argued that dopamine conveys a 'perceived saliency' signal to the ventral striatum (nucleus accumbens). Perceived saliency is formally defined as the product of stimulus saliency (a physical property of the stimulus such as its intensity relative to the background) and attentional value (a cognitive property that depends on stimulus novelty). Because perceived saliency can be higher for both appetitive and aversive stimuli, this account can explain dopamine responses to both stimulus types. Dopamine release in the ventral striatum also shows a transient increase when an expected aversive stimulus is omitted[91]. One might be tempted to reason that this is a positive RPE evoked by a 'safety' outcome, but this would be incorrect because an even larger response is evoked by the expected aversive stimulus itself. This finding also rules out an 'unsigned RPE' hypothesis, under the assumption that receipt and omission responses have similar magnitudes. By contrast, the perceived saliency hypothesis can capture this finding by virtue of the fact that the absence of a predicted stimulus (regardless of its valence) is perceived as new, thereby driving an increase in attentional value. Although this study poses a strong challenge to the RPE hypothesis, one complicating factor is that the training history of the animals may have affected the generalization of value across cues (see our earlier discussion of novelty). Furthermore, the dopamine signals were mainly recorded from the medial portion of the ventral striatum. The prevalence of perceived saliency signals in other parts of the striatum remains to be examined.

Perceived saliency can also explain why dopamine initially responds to new neutral stimuli, why this response declines across repeated presentations and why this form of stimulus pre-exposure retards subsequent stimulus-outcome learning (the latent inhibition effect)[92]. Although some aspects of these findings might be explained by the hypothesis that RPEs are normalized by uncertainty[93,94], which decreases over the course of stimulus pre-exposure, in combination with the hypothesis that dopamine encodes a novelty bonus[35], there are other aspects that are more difficult to explain in this way. Optogenetic manipulations of dopamine during pre-exposure bidirectionally affect subsequent learning, accelerating learning when excited and retarding learning when inhibited. Notably, these manipulations do not affect the conditioned response on the first learning trial, indicating that initial outcome expectations were the same across conditions. By contrast, a basic prediction of the RPE hypothesis (either normalized or unnormalized) is that increases in RPE during stimulus presentation should increase later outcome expectations.

Another alternative hypothesis is that dopamine signals retrospective inferences about 'meaningful' causal targets—stimuli that are likely causes of reward outcomes[95]. This hypothesis can explain many phenomena that were previously explained by the RPE hypothesis, such as dopamine responses to reward magnitude and probability, blocking, unblocking, overexpectation and conditioned inhibition. It can also explain new phenomena that appear to be inconsistent with the RPE hypothesis. For example, the RPE hypothesis predicts that

https://doi.org/10.1038/s41593-024-01705-4



Generalized value function: $G_t = \mathbb{E}_\pi(g_t + \gamma g_{t+1} + \gamma^2 g_{t+2} + \cdots)$

Error-driven update: $\Delta\hat{G}_t = \alpha\delta_t$

Generalized prediction error: $\delta_t = g_t + \gamma\hat{G}_{t+1} - \hat{G}_t$

Fig. 4 | A generalized prediction error framework for integrating the functional diversity of dopamine. A generalized prediction error is used to learn a 'generalized value function' that estimates the expected discounted future accumulation of a quantity g_t (the cumulant) under a policy π. Different midbrain dopamine populations (or distributed population codes) report generalized prediction errors for different outcomes, which in turn drive learning in the striatum and cortex. Different generalized prediction errors may serve different downstream functions, including guiding causal inference, identifying salient stimuli, modulating the learning rate for those stimuli and updating world models.

uncued rewards should initially evoke a large dopamine response that decreases across repeated rewards due to the acquisition of a context-outcome association. By contrast, a study in ref. 95 reports that the dopamine release in the ventral striatum actually increases across repeated rewards. The causal inference hypothesis explains this finding as the result of learning that the local reward rate (at the time of reward receipt) is higher than the background reward rate. Dopamine is hypothesized to reflect the difference between these two rates, a key computation in the identification of meaningful causal targets. Although this is an intriguing hypothesis, more modeling and experimental work is needed to assess whether the evidence truly falsifies the RPE hypothesis. Recent work has challenged some of the claims made by the retrospective inference account[96].

A third alternative hypothesis is that dopamine sets an adaptive learning rate. A study in ref. 97 has shown that licking behavior in a trace conditioning paradigm is well-explained by an RL model that adapts its learning rate based on the strength of learned sensory weights and the change in the behavioral policy at the time of reward. Intuitively, the adjustment of the policy parameters is greater when sensory information and changes in the current policy yield reward (and hence are likely to be useful for performance optimization). This model makes a distinction between 'preparatory' behavior that controls the delivery of rewards and 'reactive' behavior that controls the collection of rewards after their delivery. These two forms of behavior are both optimized for performance and linked by the same adaptive learning rate, which is putatively reported by mesolimbic dopamine. This model can capture the emergence of differences in ventral striatal dopamine levels in response to expected and unexpected rewards late in training (a canonical signature of RPEs), but, critically, it can also capture the dissociable trajectories of cue and reward responses early in training, which appear to conflict with the predictions of the RPE hypothesis[65]. Only a few studies have tested these predictions early during training, and the results are mixed. One recent study reported dopamine dynamics that conform to the predictions of the RPE hypothesis; namely, a gradual translocation of the dopamine response from rewards to cues[30].

This brief survey of alternative theories suggests that a complete theory of dopamine will require computational concepts that go beyond RPEs. Although no model exists that can capture the diverse range of empirical phenomena, in Fig. 4, we sketch what an integrative theory might look like.

Exploiting new methods to test theories of dopamine

This perspective has focused on a theory-driven approach to understanding dopamine. At the same time, testing existing theories and developing new ones depends strongly on our ability to measure relevant brain signals and extract information from them. Several technological and analytical advances have already yielded discoveries with important implications for computational theories of dopamine.

Real-time monitoring of protein kinase A (PKA) in striatal spiny projection neurons has revealed differential sensitivity to increases and decreases of dopamine in neurons that express D1Rs and D2Rs, respectively[98]. This finding provided a direct link between plasticity-related signaling and RL algorithms that apply dichotomous update rules for positive and negative prediction errors[86]. Notably, it would have been impossible to establish this link without the invention of fluorescent indicators for dynamically measuring PKA[99]. However, these findings introduce new mysteries, as phasic changes in DA concentration alter PKA signaling of >10 s, raising the question of how such prolonged changes can be useful for temporally precise credit assignment.

Another example is the study of spontaneous natural behaviors, which has been enabled by the recent development of tools for extracting low-dimensional descriptions of high-dimensional movement patterns. One such algorithm, MoSeq, extracts a time series of behavioral

'syllables' corresponding to short movement sequences such as turning, rearing and darting[100]. To understand how structured naturalistic behavioral policies emerge, Markowitz et al. used MoSeq in combination with photometric recordings of striatal dopamine transmission and closed-loop optogenetic stimulation of dopamine neuron axons[68]. Dopamine transmission had multiple effects on syllables: (1) control of syllable stochasticity, (2) reinforcement of syllable expression and (3) modulation of syllable speed. These discoveries would not have been possible without the availability of data-driven methods for analyzing naturalistic behavior.

Conclusions

Explaining the function of dopamine requires a mosaic of different mechanisms operating in different brain systems, at different timescales and in different contexts. Nonetheless, we have argued that a relatively small set of computational principles can explain a remarkably wide range of empirical facts about dopamine. Notably, this synthesis revises some fundamental concepts. Prediction errors are not limited to rewards; they sometimes need to be corrected (via value decay), and they have a central role not only in learning but also in action selection, motivation and vigor. We have also pointed out ways in which this synthesis falls short—dopamine may also signal or control salience, causal inferences, precision and learning rate. An integrative theory should incorporate these diverse mechanisms.

References

1. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
2. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
3. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141 (2005).
4. Eshel, N. et al. Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525**, 243–246 (2015).
5. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).
6. Steinberg, E. E. et al. A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).
7. Chang, C. Y. et al. Brief optogenetic inhibition of dopamine neurons mimics endogenous negative reward prediction errors. *Nat. Neurosci.* **19**, 111–116 (2016).
8. Reynolds, J. N. J. & Wickens, J. R. Dopamine-dependent plasticity of corticostriatal synapses. *Neural Netw.* **15**, 507–521 (2002).
9. Morita, K., Morishima, M., Sakai, K. & Kawaguchi, Y. Reinforcement learning: computing the temporal difference of values via distinct corticostriatal pathways: (Trends in Neurosciences 35, 457–467; 2012). *Trends Neurosci.* **40**, 453 (2017).
10. Watabe-Uchida, M., Eshel, N. & Uchida, N. Neural circuitry of reward prediction error. *Annu. Rev. Neurosci.* **40**, 373–394 (2017).
11. Starkweather, C. K. & Uchida, N. Dopamine signals as temporal difference errors: recent advances. *Curr. Opin. Neurobiol.* **67**, 95–105 (2021).
12. Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E. M. & Graybiel, A. M. Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**, 575–579 (2013).
13. Niv, Y. Neuroscience: dopamine ramps up. *Nature* **500**, 533–535 (2013).
14. Berke, J. D. What does dopamine mean? *Nat. Neurosci.* **21**, 787–793 (2018).
15. Hamid, A. A. et al. Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
16. Mohebi, A. et al. Publisher correction: dissociable dopamine dynamics for learning and motivation. *Nature* **571**, E3 (2019).
17. Krausz, T. A., Comrie, A. E., Frank, L. M., Daw, N. D. & Berke, J. D. Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron* **111**, 3465–3478 (2023).
18. Hamilos, A. E. et al. Slowly evolving dopaminergic activity modulates the moment-to-moment probability of reward-related self-timed movements. *eLife* **10**, e62583 (2021).
19. Collins, A. L. et al. Dynamic mesolimbic dopamine signaling during action sequence learning and expectation violation. *Sci. Rep.* **6**, 20231 (2016).
20. Gershman, S. J. Dopamine ramps are a consequence of reward prediction errors. *Neural Comput.* **26**, 467–471 (2014).
21. Kim, H. R. et al. A unified framework for dopamine signals across timescales. *Cell* **183**, 1600–1616 (2020).
22. Mikhael, J. G., Kim, H. R., Uchida, N. & Gershman, S. J. The role of state uncertainty in the dynamics of dopamine. *Curr. Biol.* **32**, 1077–1087 (2022).
23. Kato, A. & Morita, K. Forgetting in reinforcement learning links sustained dopamine signals to motivation. *PLoS Comput. Biol.* **12**, e1005145 (2016).
24. Beron, C. C., Neufeld, S. Q., Linderman, S. W. & Sabatini, B. L. Mice exhibit stochastic and efficient action switching during probabilistic decision making. *Proc. Natl Acad. Sci. USA* **119**, e2113961119 (2022).
25. Niv, Y. et al. Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* **35**, 8145–8157 (2015).
26. Ito, M. & Doya, K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *J. Neurosci.* **29**, 9861–9874 (2009).
27. Lloyd, K. & Dayan, P. Tamping ramping: algorithmic, implementational, and computational explanations of phasic dopamine signals in the accumbens. *PLoS Comput. Biol.* **11**, e1004622 (2015).
28. Hamid, A. A., Frank, M. J. & Moore, C. I. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* **184**, 2733–2749 (2021).
29. Guru, A., Seo, C., Kullakanda, D. S., Schaffer, J. A. & Warden, M. R. Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.21.108886 (2020).
30. Amo, R. et al. A gradual temporal shift of dopamine responses mirrors the progression of temporal difference error in machine learning. *Nat. Neurosci.* **25**, 1082–1092 (2022).
31. Ljungberg, T., Apicella, P. & Schultz, W. Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* **67**, 145–163 (1992).
32. Horvitz, J. C. Mesolimbocortical and nigrostriatal dopamine responses to salient non-reward events. *Neuroscience* **96**, 651–656 (2000).
33. Steinfels, G. F., Heym, J., Strecker, R. E. & Jacobs, B. L. Behavioral correlates of dopaminergic unit activity in freely moving cats. *Brain Res.* **258**, 217–228 (1983).
34. Gonzalez, L. S. et al. Ventral striatum dopamine release encodes unique properties of visual stimuli in mice. *eLife* **12**, e85064 (2023).
35. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).
36. 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).
37. Akiti, K. et al. Striatal dopamine explains novelty-induced behavioral dynamics and individual variability in threat prediction. *Neuron* **110**, 3789–3804 (2022).

38. Morrens, J., Aydin, Ç., Janse van Rensburg, A., Esquivelzeta Rabell, J. & Haesler, S. Cue-evoked dopamine promotes conditioned responding during learning. *Neuron* **106**, 142–153 (2020).
39. Gershman, S. J. & Niv, Y. Novelty and inductive generalization in human reinforcement learning. *Top. Cogn. Sci.* **7**, 391–415 (2015).
40. Kobayashi, S. & Schultz, W. Reward contexts extend dopamine signals to unrewarded stimuli. *Curr. Biol.* **24**, 56–62 (2014).
41. Bromberg-Martin, E. S. & Hikosaka, O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* **63**, 119–126 (2009).
42. Iigaya, K., Story, G. W., Kurth-Nelson, Z., Dolan, R. J. & Dayan, P. The modulation of savouring by prediction error and its effects on choice. *eLife* **5**, e13747 (2016).
43. Ogasawara, T. et al. A primate temporal cortex-zona incerta pathway for novelty seeking. *Nat. Neurosci.* **25**, 50–60 (2022).
44. Sharpe, M. J. et al. Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).
45. Chang, C. Y., Gardner, M., Di Tillio, M. G. & Schoenbaum, G. Optogenetic blockade of dopamine transients prevents learning induced by changes in reward features. *Curr. Biol.* **27**, 3480–3486 (2017).
46. Takahashi, Y. K. et al. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron* **95**, 1395–1405 (2017).
47. Langdon, A. J., Sharpe, M. J., Schoenbaum, G. & Niv, Y. Model-based predictions for dopamine. *Curr. Opin. Neurobiol.* **49**, 1–7 (2018).
48. Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc. Biol. Sci.* **285**, 20181645 (2018).
49. Dayan, P. Improving generalization for temporal difference learning: the successor representation. *Neural Comput.* **5**, 613–624 (1993).
50. Gershman, S. J. The successor representation: its computational logic and neural substrates. *J. Neurosci.* **38**, 7193–7200 (2018).
51. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
52. Engelhard, B. et al. Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).
53. Solié, C., Girard, B., Righetti, B., Tapparel, M. & Bellone, C. VTA dopamine neuron activity encodes social interaction and promotes reinforcement learning through social prediction error. *Nat. Neurosci.* **25**, 86–97 (2022).
54. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).
55. Cai, L. X. Distinct signals in medial and lateral VTA dopamine neurons modulate fear extinction at different times. *eLife* **9**, e54936 (2020).
56. Parker, N. F. et al. Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nat. Neurosci.* **19**, 845–854 (2016).
57. Menegas, W. et al. Dopamine neurons projecting to the posterior striatum form an anatomically distinct subclass. *eLife* **4**, e10032 (2015).
58. Menegas, W., Akiti, K., Amo, R., Uchida, N. & Watabe-Uchida, M. Dopamine neurons projecting to the posterior striatum reinforce avoidance of threatening stimuli. *Nat. Neurosci.* **21**, 1421–1430 (2018).
59. Tsutsui-Kimura, I. et al. Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. *eLife* **9**, e62390 (2020).
60. De Jong, J. W. et al. A neural circuit mechanism for encoding aversive stimuli in the mesolimbic dopamine system. *Neuron* **101**, 133–151 (2019).
61. Lammel, S., Ion, D. I., Roeper, J. & Malenka, R. C. Projection-specific modulation of dopamine neuron synapses by aversive and rewarding stimuli. *Neuron* **70**, 855–862 (2011).
62. Stalnaker, T. A. et al. Dopamine neuron ensembles signal the content of sensory prediction errors. *eLife* **8**, e49315 (2019).
63. Lee, R. S., Engelhard, B., Witten, I. B. & Daw, N. D. A vector reward prediction error model explains dopaminergic heterogeneity. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.28.482379 (2022).
64. Da Silva, J. A., Tecuapetla, F., Paixão, V. & Costa, R. M. Dopamine neuron activity before action initiation gates and invigorates future movements. *Nature* **554**, 244–248 (2018).
65. Coddington, L. T. & Dudman, J. T. The timing of action determines reward prediction signals in identified midbrain dopamine neurons. *Nat. Neurosci.* **21**, 1563–1573 (2018).
66. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016).
67. Lee, R. S., Mattar, M. G., Parker, N. F., Witten, I. B. & Daw, N. D. Reward prediction error does not explain movement selectivity in DMS-projecting dopamine neurons. *eLife* **8**, e42992 (2019).
68. Markowitz, J. E. et al. Spontaneous behaviour is structured by reinforcement without explicit reward. *Nature* **614**, 108–117 (2023).
69. Exley, R. & Cragg, S. J. Presynaptic nicotinic receptors: a dynamic and diverse cholinergic filter of striatal dopamine neurotransmission. *Br. J. Pharmacol.* **153**, S283–S297 (2008).
70. Bamford, N. S., Wightman, R. M. & Sulzer, D. Dopamine's effects on corticostriatal synapses during reward-based behaviors. *Neuron* **97**, 494–510 (2018).
71. Lindsey, J. & Litwin-Kumar, A. Action-modulated midbrain dopamine activity arises from distributed control policies. *Adv. Neural Inform. Process. Syst.* **35**, 5535–5548 (2022).
72. Greenstreet, F. et al. Action prediction error: a value-free dopaminergic teaching signal that drives stable learning. Preprint at *bioRxiv* https://doi.org/10.1101/2022.09.12.507572 (2022).
73. Howard, C. D., Li, H., Geddes, C. E. & Jin, X. Dynamic nigrostriatal dopamine biases action selection. *Neuron* **93**, 1436–1450 (2017).
74. Fry, B. R., Pence, N. T., McLocklin, A. & Johnson, A. W. Disruptions in effort-based decision-making following acute optogenetic stimulation of ventral tegmental area dopamine cells. *Learning Memory* **28**, 104–108 (2021).
75. Bova, A. et al. Precisely timed dopamine signals establish distinct kinematic representations of skilled movements. *eLife* **9**, e61591 (2020).
76. Lee, K. et al. Temporally restricted dopaminergic control of reward-conditioned movements. *Nat. Neurosci.* **23**, 209–216 (2020).
77. Eshel, N. et al. Striatal dopamine integrates cost, benefit, and motivation. *Neuron* **112**, 500–514 (2024).
78. Möller, M. & Bogacz, R. Learning the payoffs and costs of actions. *PLoS Comput. Biol.* **15**, e1006285 (2019).
79. Van Swieten, M. M. H. & Bogacz, R. Modeling the effects of motivation on choice and learning in the basal ganglia. *PLoS Comput. Biol.* **16**, e1007465 (2020).
80. Wan, Y., Naik, A. & Sutton, R. S. Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, 10653–10662 (PMLR, 2021).
81. Daw, N. D. & Touretzky, D. S. Long-term reward prediction in TD models of the dopamine system. *Neural Comput.* **14**, 2567–2583 (2002).

82. Aberman, J. E. & Salamone, J. D. Nucleus accumbens dopamine depletions make rats more sensitive to high ratio requirements but do not impair primary food reinforcement. *Neuroscience* **92**, 545–552 (1999).
83. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507–520 (2007).
84. Berridge, K. C. & Robinson, T. E. What is the role of dopamine in reward: hedonic impact, reward learning, or incentive salience? *Brain Res. Brain Res. Rev.* **28**, 309–369 (1998).
85. Sippy, T. & Tritsch, N. X. Unraveling the dynamics of dopamine release and its actions on target cells. *Trends Neurosci.* **46**, 228–239 (2023).
86. Collins, A. G. E. & Frank, M. J. Opponent actor learning (OpAL): modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* **121**, 337–366 (2014).
87. McClure, S. M., Daw, N. D. & Montague, P. R. A computational substrate for incentive salience. *Trends Neurosci.* **26**, 423–428 (2003).
88. Jaskir, A. & Frank, M. J. On the normative advantages of dopamine and striatal opponency for learning and choice. *eLife* **12**, e85107 (2023).
89. Bogacz, R. Dopamine role in learning and action inference. *eLife* **9**, e53262 (2020).
90. Kutlu, M. G. & Schmajuk, N. A. Solving Pavlov's puzzle: attentional, associative, and flexible configural mechanisms in classical conditioning. *Learn. Behav.* **40**, 269–291 (2012).
91. Kutlu, M. G. et al. Dopamine release in the nucleus accumbens core signals perceived saliency. *Curr. Biol.* **31**, 4748–4761 (2021).
92. Kutlu, M. G. et al. Dopamine signaling in the nucleus accumbens core mediates latent inhibition. *Nat. Neurosci.* **25**, 1071–1081 (2022).
93. Gershman, S. J. Dopamine, inference, and uncertainty. *Neural Comput.* **29**, 3311–3326 (2017).
94. Möller, M., Manohar, S. & Bogacz, R. Uncertainty-guided learning with scaled prediction errors in the basal ganglia. *PLoS Comput. Biol.* **18**, e1009816 (2022).
95. Jeong, H. et al. Mesolimbic dopamine release conveys causal associations. *Science* **378**, eabq6740 (2022).
96. Qian, L. et al. The role of prospective contingency in the control of behavior and dopamine signals during associative learning. Preprint at *bioRxiv* https://doi.org/10.1101/2024.02.05.578961 (2024).
97. Coddington, L. T., Lindo, S. E. & Dudman, J. T. Mesolimbic dopamine adapts the rate of learning from action. *Nature* **614**, 294–302 (2023).
98. Lee, S. J. et al. Cell-type-specific asynchronous modulation of PKA by dopamine in learning. *Nature* **590**, 451–456 (2021).
99. Lee, S. J., Chen, Y., Lodder, B. & Sabatini, B. L. Monitoring behaviorally induced biochemical changes using fluorescence lifetime photometry. *Front. Neurosci.* **13**, 766 (2019).
100. Wiltschko, A. B. et al. Mapping sub-second structure in mouse behavior. *Neuron* **88**, 1121–1135 (2015).
101. Hamid, A. A. Dopaminergic specializations for flexible behavioral control: linking levels of analysis and functional architectures. *Curr. Opin. Behav. Sci.* **41**, 175–184 (2021).
102. Hughes, R. N. et al. Ventral tegmental dopamine neurons control the impulse vector during motivated behavior. *Curr. Biol.* **30**, 2681–2694 (2020).
103. De Jong, J. W., Liang, Y., Verharen, J. P. H., Fraser, K. M. & Lammel, S. State and rate-of-change encoding in parallel mesoaccumbal dopamine pathways. *Nat. Neurosci.* **27**, 309–318 (2024).
104. Gerfen, C. R. & Surmeier, D. J. Modulation of striatal projection systems by dopamine. *Annu. Rev. Neurosci.* **34**, 441–466 (2011).
105. Durstewitz, D. & Seamans, J. K. The dual-state theory of prefrontal cortex dopamine function with relevance to catechol-o-methyltransferase genotypes and schizophrenia. *Biol. Psychiatry* **64**, 739–749 (2008).
106. Mahadevia, D. et al. Dopamine promotes aggression in mice via ventral tegmental area to lateral septum projections. *Nat. Commun.* **12**, 6796 (2021).
107. Dabney, W. et al. A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).
108. Lowet, A. S. et al. An opponent striatal circuit for distributional reinforcement learning. Preprint at *bioRxiv* https://doi.org/10.1101/2024.01.02.573966 (2024).
109. Grace, A. A. Phasic versus tonic dopamine release and the modulation of dopamine system responsivity: a hypothesis for the etiology of schizophrenia. *Neuroscience* **41**, 1–24 (1991).
110. Kilpatrick, M. R., Rooney, M. B., Michael, D. J. & Wightman, R. M. Extracellular dopamine dynamics in rat caudate-putamen during experimenter-delivered and intracranial self-stimulation. *Neuroscience* **96**, 697–706 (2000).

Acknowledgements

We are grateful to the many students and postdocs who have worked on aspects of this research with us. The ideas presented here also benefited from conversations with N. Daw. This work was supported by the National Institutes of Health (U19-NS-113201-01) and the Air Force Office of Scientific Research (FA9550-20-1-0413).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Samuel J. Gershman.

Peer review information *Nature Neuroscience* thanks Arif Hamid and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature America, Inc. 2024