A vector reward prediction error model explains dopaminergic heterogeneity

Rachel S. Lee[1], Ben Engelhard[1], Ilana B. Witten[1,2 *], Nathaniel D. Daw[1,2 *]

Affiliations:

[1]Princeton Neuroscience Institute
[2]Department of Psychology
Princeton University, Princeton NJ USA 08544

The hypothesis that midbrain dopamine (DA) neurons broadcast an error signal for the prediction of reward (reward prediction error, RPE) is among the great successes of computational neuroscience[1–3]. However, recent results contradict a core aspect of this theory: that the neurons uniformly convey a scalar, global signal. Instead, when animals are placed in a high-dimensional environment, DA neurons in the ventral tegmental area (VTA) display substantial heterogeneity in the features to which they respond, while also having more consistent RPE-like responses at the time of reward. Here we introduce a new "Vector RPE" model that explains these findings, by positing that DA neurons report individual RPEs for a subset of a population vector code for an animal's state (moment-to-moment situation). To investigate this claim, we train a deep reinforcement learning model on a navigation and decision-making task, and compare the Vector RPE derived from the network to population recordings from DA neurons during the same task. The Vector RPE model recapitulates the key features of the neural data: specifically, heterogeneous coding of task variables during the navigation and decision-making period, but uniform reward responses. The model also makes new predictions about the nature of the responses, which we validate. Our work provides a path to reconcile new observations of DA neuron heterogeneity with classic ideas about RPE coding, while also providing a new perspective on how the brain performs reinforcement learning in high dimensional environments.

Introduction

Among the more prominent hypotheses in computational neuroscience is that phasic responses from midbrain dopamine (DA) neurons report a reward prediction error (RPE) for learning to predict rewards and choose actions[1–3]. While clearly stylized, this account has impressive range, connecting neural substrates (the spiking of individual neurons and plasticity at target synapses, e.g. in striatum[4]) to behavior (trial-by-trial adjustments in choice tendencies[5]), all via interpretable computations over formally defined decision variables. The question of this article is whether and how the strengths of this account can be reconciled with a growing body of evidence challenging a core feature of the model, i.e. the identification of a scalar, globally broadcast RPE signal in DA responses.

This scalar RPE is not a superficial claim of these theories, but instead one that connects a key computational idea to a number of empirical observations. Computationally, scalar decision variables reflect the ultimate role of any decision system in comparison: the decision-maker must order different outcomes against one another to choose which to take. Thus, even though value arises from multiple incentives (water, food, etc.), these must effectively be reduced to a so-called "common currency" for comparison[6]. In turn, the error in these value predictions (e.g. the difference between expected and obtained overall value) is then also scalar. Such scalar comparisons have been argued to be apparent in features of Pavlovian conditioning such as transreinforcer blocking[7,8], by which different good or bad outcomes can substitute for one another. Neurally, the scalar nature of RPE was also historically viewed as a good fit for a number of features of the DA system. Anatomically, the ascending DA projection has an organization more consistent with a "broadcast" than a labeled line code: a relatively small number of individual neurons innervate a large area of the forebrain via diffuse projections[9]. For instance, individual DA neurons branch extensively to innervate large areas (~1 mm^3) of striatum[10], where volumetric propagation of released DA to extrasynaptic DA receptors further blurs its effect[11]. Physiologically, early reports also stressed the homogeneity of responses of midbrain DA neurons on simple conditioning tasks – e.g., a large majority of units respond to unexpected reward[12].

However, the physiological argument for a scalar RPE is increasingly untenable, as a mounting body of recent work challenges the generality of this finding by demonstrating a range of variation in dopamine responses. In particular, midbrain DA neurons can have heterogeneous and specialized responses to task variables during complex behavior[5,13–28], even while having relatively homogenous responses to reward[3,15,21,29,30]. We investigate these phenomena by focusing on a recent study from our labs which provides one of the most detailed and dramatic examples of this pattern, at the level of single neurons. By performing 2-photon imaging across a population of VTA DA neurons while mice performed an evidence accumulation task in a virtual reality T-maze, we observed that while neurons respond relatively homogeneously to reward during the outcome period, during the navigation and decision period, they respond heterogeneously to kinematics, position, cues, and more[15].

How can we reconcile such DAergic heterogeneity with the substantial evidence for the RPE theory? Here, we show that these properties will emerge once we address two key oversimplifications of the classic RPE model: one anatomical and one computational.

Anatomically, although the ascending DAergic projection is relatively diffuse[9,31], inputs to DA neurons are not homogenous but instead arise from cortico-basal-ganglionic circuits that are highly topographically organized[32–34]. Computationally, although RPE is a function of value (which, as usually defined, is scalar), value is itself a function of a variable known in RL models as "state": a summary of the current situation in the task that is supposed to reflect all information that is relevant to reward prediction and choice. Although theoretical simulations have often employed simplified "grandmother cell" codes for state[1–3], in a realistic environment or a biological brain, these stand in for a high-dimensional group of sensory and internal variables that are likely widely distributed throughout the brain. Here we propose that a distributed code for state is carried by corticostriatal circuits, which transform it into corresponding distributed codes for value and RPE that, in effect, decompose these scalar variables over state features[34]. In this way, different striatal and DAergic neurons reflect the contribution of different state features to value and RPE, but the ensemble collectively represents canonical RL computations over the scalar variables.

Our new model offers a number of key insights. First, it explains the striking contrast observed in dopamine neurons: heterogeneous responses to task variables alongside much more uniform outcome-period responses[15,21]. This is a key empirical signature of the distributed value code we posit, because responses in the navigation and decision period arise from value predictions (and hence ultimately from diverse state features) whereas the outcome-period response is driven by common reward information. Second, the model retains an algebraic mapping to the standard theory[1–3], preserving its successes while improving its match to anatomical and physiological evidence; it also lays a foundation for further elaborations that take advantage of the distributed code to improve learning. Finally, the theory exposes an unexpected connection between the puzzling empirical phenomena of DAergic heterogeneity and a major theoretical question in RL models in neuroscience: the nature of state. While there has been substantial theoretical interest in the principles by which the brain represents task state[35–38], there is relatively little empirical evidence to constrain these ideas. The new model suggests that the DA population representation itself can provide a window into this hitherto elusive concept of the neural representation of state.

Results

The Vector RPE model

Here, we propose a Vector RPE model as an extension of the classic scalar RPE model (**Fig. 1**). RL models typically assume that the goal of the learner is to learn the value function $V(s_t) = E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid s_t\right]$ (i.e. the expected sum of γ-discounted future rewards r_t starting in some state s_t). Both in neuroscience and in AI[39], a typical starting assumption for high-dimensional or continuous tasks is to assume the learner approximates value linearly in some feature basis. That is, it represents the state s_t by a vector of features $\vec{\phi}(s_t)$ (hitherto, $\vec{\phi}_t$) and approximates value as a weighted sum of those features, i.e. $V_t = V(s_t) \approx \vec{\phi}_t \cdot \vec{w}$. This

reduces the problem of value learning (for some feature set) to learning the error-minimizing weights $\vec{w}$, and more importantly for us, formalizes the state representation itself as a vector of time-varying features $\vec{\phi}_t$. (The linearity assumption is less restrictive than it seems, since the features may be arbitrarily complex and nonlinear in their inputs. For instance, this scheme is standard in AI, including at the final layer in deep-RL models, which first derive features from a video input by multilayer convolutional networks, then finally estimate value linearly from them[40,41].)

In a standard temporal-difference (TD) learning model, weights are learned by a delta rule using the RPE $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$. A typical cartoon of how these are mapped onto brain circuitry is shown in **Fig. 1a**, with a cortical input population vector for state features projected to scalar value and RPE stages, corresponding, respectively, to presumed uniform populations of striatal and midbrain DA neurons[1–3]. The RPE then drives weight learning at the corticostriatal synapses via ascending DA projections.

We propose to relax the unrealistic assumption that the corticostriatal stage of this circuit involves complete, uniform convergence from vector state to scalar value (**Fig. 1b**). In fact, projections are substantially topographic at each level of the corticostriatal circuit[31,32,42]. Thus, if different striatal units i preferentially receive input from particular cortical features $\phi_{i,t}$, then value will itself be represented by a distributed feature code $V_{i,t} = w_i\phi_{i,t}$ (see also[2,43]), and (in turn) DA neurons preferentially driven by each "channel" will compute feature-specific prediction errors

$$\delta_{i,t} = \frac{r_t}{N} + \gamma V_{i,t+1} - V_{i,t} = r_t + w_i(\gamma\phi_{i,t+1} - \phi_{i,t}) \quad [1]$$

(where N is a scale factor equaling the number of channels). Importantly, due to linearity, the aggregate response (summed over channels i) at the value stage reflects the original scalar value, and the aggregate response at the RPE stage corresponds to the original scalar RPE.

Thus (assuming the ascending dopaminergic projection is sufficiently diffuse as to mix the channels prior to the weight update) the model corresponds algebraically to the classic one, just mapped onto the brain circuitry in a more realistic way. Note that our basic insight that nonuniform projections will result in value and RPE stages that reflect input feature variation, but preserve the model's function due to linearity, holds under still more realistic models in which channels partially mix at each step due to anatomically delimited convergence, or in which linearity is only approximate, etc. Also, the assumption of perfect mixing on the ascending DA stage is needed only to recover the classic model exactly. The current framework also admits variants which maintain separation on the ascending signal. They function similarly, but allow for more efficient, "divide and conquer" learning in some situations[44](see Discussion).

Anatomically, this account (while still exceptionally stylized) represents a more realistic picture compared to the traditional scalar story of both value and RPE. Overall, there is clearly topography preserved in each stage of projection, from cortical inputs to MSNs, from MSNs to dopaminergic units, and indeed from dopaminergic units back to forebrain[32–34]. Each of these

stages does involve convergence or dimensionality reduction (and the model could accommodate any degree of summation at each stage), but it seems plausible that the most drastic such convergence is the projection from dopamine back to striatum. Indeed each dopamine axon entering striatum branches into an extensive, dense arborization which synapses on many MSNs over a sizable territory of striatum[10]. Moreover, dopamine is also released at nonsynaptic release sites and propagates volumetrically to nonsynaptic receptors[11].

Physiologically, this model is constructed so that (even without specifying a particular feature basis) it captures several aspects of the heterogeneous dopamine response. Consider times when primary reward is not present, such as during the navigation and decision-making period as in[15]. Here, $r_t = 0$ and Equation 1 reduces to $\delta_{i,t} = w_i(\gamma\phi_{i,t+1} - \phi_{i,t})$: that is, each DA unit reports the time-differenced activity in its feature, weighted by its own association w_i with value. Depending on what the features $\vec{\phi}_t$ are, this would explain dopaminergic response correlations with different, arbitrary covariates: to the extent some feature ϕ_i is task-relevant, it will have nonzero w_i (where the sign of w_i is determined by ϕ_i's partial correlation with value in the presence of the other features), and its derivative will correlate with a subset of neurons. Even a DA neuron that is driven by an objectively task-irrelevant feature is likely to respond to it, due to incidental or transient correlations between the feature and value producing nonzero w_i.

Conversely, at outcome time, all modeled RPE units are likely to respond differentially for reward than nonreward (due to r_t being shared across all channels in Equation 1) as in[15]. This reward response may also be modulated by its predictability, due to each channel's share of the temporal difference component, $w_i(\gamma\phi_{i,t+1} - \phi_{i,t})$, but is unlikely to be completely predicted away by most individual features. Finally, since the standard RPE δ_t is equal to the sum over all channels $\sum_i \delta_{i,t}$ by construction, the model explains why neuron-averaged data (or bulk signals as in fiber photometry or BOLD), as often reported, resemble TD model predictions even potentially in the presence of much inter-neuron variation.

Deep RL network to simulate the Vector RPE model

Although our Vector RPE model is fully general, in order to simulate the model in the context of a specific task, we need to specify an appropriate set of basis functions for the task state. We consider our previously reported experiment in which mice performed an evidence accumulation task in a virtual reality environment while VTA DA neurons were imaged [15] (**Fig. 2a**). In this task, mice navigated in a virtual T-maze while viewing towers that appeared transiently to left and right, and were rewarded for turning to the side where there had been more towers.

To simulate the vector of RPEs in this task (which we will simply refer to as the "vector RPE"), we took advantage of the fact that this task was based in virtual reality, and therefore we could train a deep RL agent on the same task that the mice performed to derive a vector of features,

and in turn, a corresponding distributed code for values and RPEs (**Fig. 2b**). In particular, we used a deep neural network to map the visual images from the virtual reality task to 64 feature units (via three convolutional layers for vision, then a layer of LSTM recurrent units for evidence accumulation; see Methods). These features were then used as common input for an actor-critic RL agent: a linear value-predicting critic (as above, producing the vector RPE that sums to the traditional scalar RPE) and a softmax policy learner responsible for choosing an action (left, right, forward) at each step. We trained the network to perform the task using the A2C algorithm[41].

After training, the agent accumulated evidence along the central stem of the maze to ultimately choose the correct side with accuracy similar to mice (shown as a psychometric curve in **Fig. 2c**). A minimal abstract state space underlying the task is 2D, consisting of the position along the maze and the number of towers seen (on left minus right) so far. When examining average responses of the trained state features from the network in this space, we find that units are tuned to different combinations of these features, implying that they collectively span a relevant state representation for the task (**Extended Data Fig 1**). Further, the scalar value function output by the trained agent while traversing the maze, derived by summing the value vector, is modulated by trial difficulty (operationalized, following the mouse study, as the absolute value of the difference in the number of cues presented on either side), meaning that the trained agent can predict the likelihood of reward on each trial (**Fig. 2d**).

Vector RPE has heterogeneous selectivity during the cue period

The key finding from our prior cellular resolution imaging of DA neurons during the virtual T-maze task was heterogenous coding of task and behavioral variables during the navigation and decision period, such as view angle, position, and cue side (contralateral versus ipsilateral)[15]. This heterogeneity was followed by relatively homogeneous responses to reward during the outcome period.

We first sought to determine if the vector RPEs from our agent had heterogenous tuning to various behavioral and task variables during the cue period, similar to our neural data. We first considered the view angle during the central stem of the maze, which had no effect on reward delivery in the simulated task, but which the agent could nonetheless rotate by choosing left or right actions while in the stem of the maze. The vector RPE displayed idiosyncratic selectivity across units for the range of possible view angles (**Fig. 3a**), which qualitatively resembled our previous results from DA neuron recordings (**Fig. 3d**). We next considered position tuning along the central stem of the maze. A subset of RPE units showed position selectivity, including both downward and upward ramps towards the end of the maze (**Fig. 3b**), again qualitatively resembling our neural recordings (**Fig. 3e**). Finally, units also had idiosyncratic and heterogeneous cue selectivity, including preference for right vs left cues and diversity in the timing of the response (**Fig. 3c**). This was reminiscent of the side-selectivity of the *in vivo* cue responses, although the time courses of the artificial agent RPEs and *in vivo* calcium indicator data differed (**Fig. 3f**), presumably due to temporal filtering in the latter.

Reward-irrelevant features are present in the Vector RPE

A feature of the neural data is the encoding of task features that appear to be reward-irrelevant[15]. In the model, the requirement of the network to extract relevant task state from the high-dimensional video input implies the possibility that reward-irrelevant aspects of the input may "leak" into the state features, and ultimately into the vector RPE – even if they average out in the scalar RPE. Although we of course do not intend backpropagation as a mechanistic account of how the brain learns features, its use here exemplifies the more general problem that a low-dimensional output objective (choosing actions and predicting scalar reward) imposes few constraints on higher-dimensional upstream feature representations.

To test the validity of this idea, we sought to examine whether there may be coding of reward-irrelevant visual information in the Vector RPE of the agent. We focused on unambiguously reward-irrelevant visual structure in the task, namely incidental background patterns that appear on the wall in the stem of the maze (**Fig. 4a**). This background pattern repeats every 43 cm, a structure which is clearly visible as off-diagonal banding in the matrix of similarity between pairs of video frames across all combinations of locations. The structure is also visible as peaks in the 1D function (similar to an autocorrelation) showing the average similarity as a function of the distance between frames (**Fig. 4b**). To investigate whether the same irrelevant feature dimensions are present at the level of the vector RPEs, we repeat the same analysis on them. (To ensure the network inputs actually reflect such repeating similarity structure, for this analysis we exposed the trained network to a maze traversal with a fixed view angle and no cue towers.) The resulting vector RPEs show the same pattern of enhanced similarity at the characteristic 43 cm lag, supporting the expectation that task-irrelevant features do propagate through the network (**Fig. 4c-d**). This particular effect remains a prediction for future neural experiments: because the fixed view-angle condition was not run in the mouse studies, we cannot repeat this test in existing empirical data.

Cue responses are consistent with feature-specific RPEs

While our Vector RPE model implies idiosyncratic and even task-irrelevant tuning in individual DA neurons, it also makes a fundamental prediction about the nature of these responses. In particular, responses to individual features represent not generic sensory or motor responses but feature-specific components of RPE. What in practice this means for a given unit depends both on what is its input feature $\vec{\phi}_t$, and also what other features are represented. But in general we would expect that units that appear to respond to some feature (such as contralateral cues) do not reflect simple sensory responses (the presence of the cue) but rather should be further modulated by the component of RPE elicited by the feature. This could be particularly evident when considering the response averaged over units selective for a feature.

To test this hypothesis, we performed a new analysis, by subdividing cue-related responses (which are largely side-selective in both the model and neural data, **Fig. 5**) to determine if they were, in fact, additionally sensitive to the prediction error associated with a cue on the preferred side. For this, we distinguished these cues as *confirmatory* – those that appear when their side has already had more cues than the other and therefore (due to the monotonic psychometric

curve, **Fig. 2c**) are associated with an increase in the probability that the final choice will be correct and rewarded, i.e. positive RPE – vs *disconfirmatory* – cues whose side has had fewer towers so far and therefore imply a decreased probability of reward (**Fig. 5a**). As expected, when we considered the population of cue-onset responding vector RPE units from the deep RL agent, these responses were stronger, on average, for confirmatory than disconfirmatory cues (**Fig. 5b**), reflecting the component of RPE associated with the cue. We next reanalyzed our previous DA recordings based on this same insight. Consistent with the hypothesis that these cue responses reflect partial RPEs for those cues, the responses of cue-selective DA neurons were much stronger for confirmatory than disconfirmatory cues (**Fig. 5c**). This implies that the heterogenous cue selectivity in DA neurons is indeed consistent with a cue-specific RPE. Importantly, the fact that these cue-responsive neurons are overwhelmingly selective for contralateral cues implies that these responses, combined across hemispheres, simultaneously represent separate components of a 2-D vector RPE.

Uniform responses to reward at outcome period

In addition to explaining heterogeneity during the navigation and decision-making period, the Vector RPE model also explained the contrasting homogeneity of the neural responses to reward during the outcome period. Reflecting the standard properties of an RPE, the simulated scalar RPE (averaged over units) responded more for rewarded than unrewarded trials (**Fig. 6a**). Since this aspect of the response ultimately arises (in Equation 1) from a scalar reward input, it is highly consistent across the units (**Fig. 6b**), matching the neural data from our experiment (**Fig. 6c**) and the widely reported reward sensitivity of DAergic units (N = 303).

Equation 1 also implies a subtler prediction about the vector RPE, which is that although the modulation by reward is largely uniform across units, the simultaneous modulation of this outcome response by the reward's predictability should be much more variable. This is because this latter modulation arises from the value terms in Equation 1, which are distributed across features (i.e., value is a vector). In this task, reward expectation can be operationalized based on the absolute difference in tower counts, which is a measure of trial difficulty (predicting the actual chance of success as shown in **Fig. 2c**), That is, when a reward occurs following a more difficult discrimination, the agent will have expected that reward with lower probability (**Fig. 2d**). Accordingly, the simulated scalar RPE was larger for rewards on hard than easy trials (**Fig. 6d**; **Extended Data Fig. 2**). However, when broken down unit-by-unit, although the median of individual units in the Vector RPE was consistent with the scalar RPE ($P < 0.05$ two-sided Wilcoxon signed rank test for N=64), the size and direction of this effect varied widely across units (**Fig. 6e**). A similar finding emerged from the neural data: while on average the reward response across reward-responsive population was modulated as expected by expectation ($P < 3e\text{-}5$ two sided Wilcoxon signed rank test for N = 303), there was high variability in the direction and extent of modulation across units (**Fig. 6f**).

Discussion

Here we propose a new theory which helps to reconcile recent empirical reports of DA response heterogeneity to the classic idea of DA neurons as encoding RPEs. Our model posits that DA

heterogeneity is in part a reflection of a high-dimensional state representation, and thus the DA responses form a distributed RPE code with respect to features from the state input. We show how this model produces heterogeneous responses to task variables, but relatively uniform responses to reward, recapitulating recent empirical work[15]. We also test the model's prediction that heterogeneous DA responses are not simply responses to sensory and behavioral features of the task, but instead reflect components of the RPE with respect to a subset of the features.

Aspects of DAergic heterogeneity that can and cannot be explained by our model

The question arises of which aspects of the many experimentally reported instances of DAergic variation our model can explain, versus which may reflect additional (not necessarily mutually exclusive) mechanisms. It can be helpful to categorize empirical studies of DA into those describing heterogeneous responses at outcome (to rewards, omissions, or punishments[16,18,22,25,45–52]) versus heterogeneous responses to other task events, including stimuli and movements[5,13,15,21,22,24,26,28,51–56].

Regarding the latter, the basic insight of our model is that an arbitrary population code for state, if not fully convergent onto a population of DA neurons, can give rise to diverse patterns of simultaneous, multiplexed responses to different nonreward task events. Collectively, these responses constitute a population code over feature-specific RPEs in an otherwise standard TD learning setting aiming to predict a single, scalar reward input. In principle (given corresponding feature inputs) this architecture could explain a wide range of reports of multiplexed and idiosyncratic DA responses to different features, including both stimuli and movements and responses with different temporal patterns (including both ramping and waves)[26,57]. Although the model is extremely general in this respect, since different DA response patterns can arise from different feature inputs, it does make specific and testable predictions. For instance, the model predicts that DA encoding of task features, in general, actually reflects components of RPE with respect to each feature, rather than strictly the main effect of the feature itself. For instance, in the current dataset DA neurons distinguish confirmatory from disconfirmatory cues (**Fig 5b,c**), which differ in their reward consequences. This can be quite subtle to test, and most existing reports of apparently heterogeneous DAergic sensitivity have not addressed it.

Regarding outcomes, a hallmark of our model is that the heterogeneous responses to task variables coexist with a classic and more uniform main effect of reward versus omission (**Fig 6b,c**). The current theory can accommodate some other types of heterogeneity at outcome: for instance, differences between neurons in expectancy effects as in **Fig 6e,f**, including responses to consumption-related sensory or motor features like licking. But our account, by itself, does not explain other reports of variation across neurons or regions in the overall sensitivity to reward and punishment, or "salience"-like outcome responses. Such variation has also been clearly demonstrated[16,25,46–50,58], but most often arises when comparing responses across more spatially distant regions.

Indeed, reports of DAergic heterogeneity differ not just to the nature of the responses, but also as to the spatial scale. For instance, while some studies concern neuron-to-neuron variation within the VTA[15,21,45], others consider larger-scale variation[5,16,20,22,25–27,47–50,52,53,59], often using fiber

photometry or voltammetry. Our theory might, in principle, explain some variation in both scales, since inhomogeneous state feature input to DA neurons might vary over small and large scales. However, larger scale, inter-area variation seems more likely, at least in part, to reflect additional functional or anatomical differences beyond those contemplated by our model. For instance, different neurons in a putative VTA-NAc critic circuit might constitute a population code over state features (as in our model), but DA input to more dorsal areas of striatum might, hypothetically, support a distinct functional role (e.g., in RL terms an "actor") in controlling movement. In fact, contralateral movement sensitivity in the DMS DA signals appears to reflect the movement direction per se, and is not (as would be predicted by our model) further modulated by the RPE with respect to the movement[13]. A second point is that, since our model predicts that averaging over features will recover the original scalar RPE, the types of variation it envisions will tend to be washed out by methods like photometry that involve averaging over many neurons. Thus, in all, we view our model primarily as addressing interneuron variation at a relatively small spatial scale (e.g., within VTA or even a part of it), while not ruling out additional sources of variation, especially across regions.

Alternative computational accounts for dopaminergic heterogeneity

Most previous theories have taken a substantially different approach to explaining DAergic heterogeneity, by positing multiple distinct error signals, each specialized for learning a different target function[44,45,60–70].

For instance, a family of error signals could be used to learn to predict different outcomes such as rewards versus punishments[47,48,66,71], to learn the rewards associated with different actions or effectors[44], to predict rewards at different temporal scales[27,61], or to track different goals and subgoals in a hierarchical task[68,69,72]. What these examples have in common is that they each posit a handful of error signals, which are each associated with DA activity in spatially separate DA nuclei or target regions. This commonality is likely no accident, since the relatively diffuse ascending DAergic anatomy seems poorly suited for supporting a larger number of finer-grained closed circuits, with many different error signals training many distinct predictions at nearby targets. For the same reason, while such large-scale functional variation may represent an additional source of heterogeneity over and above the finer-scale variation our model discusses, it seems less plausible that this type of scheme can explain the diverse VTA responses we discuss here.

A related point is that many (though not all) of these previous vector-valued RPE proposals ground the different error signals in different reward or outcome signals: i.e., different RPEs are defined, each for predicting different outcomes. Apart from reward vs. punishment, examples of this approach include predicting different scaled functions of reward amount (which enables learning different quantiles of the distribution over stochastic rewards[45,67]; and, in the most extreme case, treating each possible sensory observation or state as its own separate prediction target, thus learning a very high-dimensional set of value functions for predicting many different future events (the "successor representation")[60]. A key difference from our approach is that this last model predicts that the heterogeneous navigation and decision period

responses should follow from corresponding heterogeneous outcome responses, but this contradicts the striking homogeneity of outcome responses in[15].

In some respects, our model's explanation of heterogeneity is more similar to another recent theory[70] in which heterogeneity also emerges from nonuniform anatomical connectivity rather than being imposed by top-down normative considerations. A main difference is that we frame our vector RPE model in terms of classic RL algorithms, building more closely on the TD literature and showing how it can be extended to accommodate this type of heterogeneity.

Variations, computational benefits, and additional predictions of our Vector RPE model

For simplicity, we have presented our Vector RPE model, as schematized in **Fig. 1b**, with stylized anatomical assumptions: perfect one-to-one connectivity in the descending projections between cortical state, striatal value, and midbrain RPE neurons, versus complete convergence in the ascending DA projections to striatum. However, our results do not depend on these assumptions. First, the model can accommodate any degree of convergence in the descending stages. In this case, heterogeneous cue-period responses (though blended to some degree) will still be observed at the DA layer, and the model will still implement standard TD learning in the sense that the averages over RPE and value units will recover the same scalar functions.

A more interesting variant arises from relaxing the assumption that the ascending DA projections mix perfectly. This produces a model in which different DA target regions receive different error signals, reflecting information about different state features – e.g., sensory modalities such as vision vs. audition. While this approach no longer corresponds algebraically to the standard scalar RPE model, models of this sort have a long history in animal conditioning[73,74] and can perform appropriately in many situations[75]. A main behavioral prediction of this variant is that, in conditioning experiments, cues that share an error signal will compete with one another in predicting reward and thereby show blocking effects[76,77], whereas cues processed in distinct RPE channels will not block each other. Thus, particular behavioral effects (patterns of blocking) are predicted depending on the distribution of DA cue responses and their patterns of connectivity, overlapping or separate.

This example speaks to a different question, which is whether the proposed vector RPE architecture has computational advantages relative to the classic scalar broadcast model. To be clear, a primary contribution of our model is the demonstration that heterogeneity can arise even in the classic model under more realistic anatomical assumptions. But decomposing values and RPEs according to stimuli or state features, and organizing these modules spatially, offers a number of advantages for both value prediction and learning. For prediction, this can enable individualized gain control, such as feature-selective attention[78], and other divide-and-conquer schemes to focus learning on context-relevant dimensions[44], Litwin-Kumar, personal communication). For learning, a key feature of backpropagation in artificial deep network models is per-synapse specialized training signals. While this remains biologically implausible, the decomposition of feature channels in the current model offers a biologically realizable substrate for more limited targeting.

Population codes for state

Although much RL modeling in neuroscience assumes a simple state representation that is hand-constructed by the modeler[2,3]; in general, an effective state representation depends on the task, and the brain must learn or construct it autonomously as part of solving the full RL problem. How it does this is arguably the major open question in these models. Indeed, in AI, recent progress on this problem (notably using deep neural networks) has been the main innovation fueling impressive advances scaling up otherwise standard RL algorithms to solve realistic, high-dimensional tasks like video games[40,79]. In psychology and neuroscience also, there have been a number of recent theoretical hypotheses addressing how the brain might build states, such as the successor representation and latent-state inference models[36,80–84]. But there exist relatively few experimental results to assess or constrain these ideas, for instance because learning behavior alone is relatively uninformative about state, whereas in the brain it is unclear which neural representations directly play this role for RL (e.g., grid cells vs. place cells for spatial tasks[83,85];).

A main consequence of the new model is that, if our hypothesis is correct, then the heterogeneous DAergic population itself gives a new experimental window, from the RL system's perspective, into the brain's population code over state features. While the model itself is agnostic to the feature set used, the various DA responses should in any case reflect it. This builds on previous work showing that even a scalar DAergic TD error signal can be revealing about the upstream state that drives it[35,37,38], but on the new theory, the vector DA code much more directly reflects the upstream distributed code for state. Thus, the theory offers a general framework to reason quantitatively about population codes for state. This should enable new experiments and data analyses to infer the brain's specific state representation from neural recordings and in particular to test ideas about how it is built: how it changes across different tasks and as tasks are acquired.

Acknowledgements

We thank Alvaro Luna for help with the VR software system, Michelle Lee and Erin Grant for help with training the deep RL network, Peter Dayan, Ari Kahn, and Lindsey Brown for comments on this work, and additionally the rest of the Daw and Witten labs for their help. This work was supported by an NSF GRFP (RSL), 1K99MH122657 (BE), NIH R01 DA047869 (IBW), U19 NS104648-01 (IBW), ARO W911NF-16-1-0474 (NDD), ARO W911NF1710554 (IBW), Brain Research Foundation (IBW), Simons Collaboration on the Global Brain (IBW), and the New York Stem Cell Foundation (IBW). IBW is a NYSCF—Robertson Investigator.

Figures

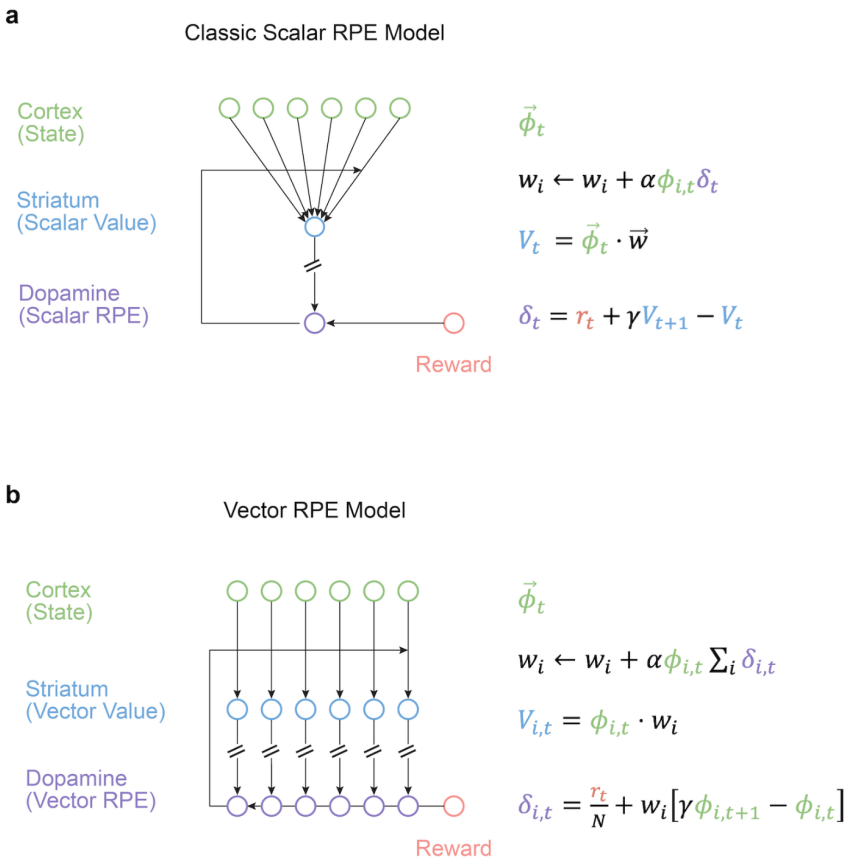


Fig. 1: Vector RPE Model updates classic TD learning to produce heterogeneous DA signals that reflect the state representation. (a) Classic mapping between equations of TD learning model and brain circuitry[1–3]. **(b)** Our proposed Vector RPE model, which remaps the same algorithm onto brain circuitry such that value and RPE are vectors but the overall computations are preserved.

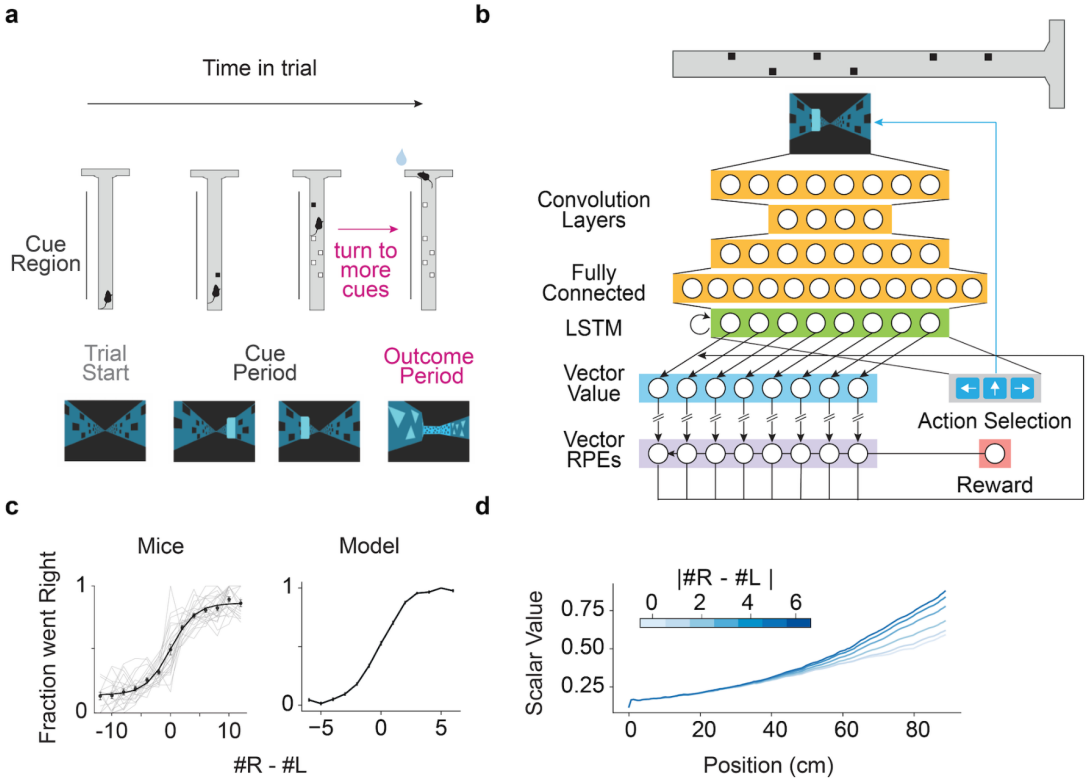


Fig. 2: Deep reinforcement learning network trained on VR evidence accumulation task from Engelhard et al.[15] **(a)** Task schematic of the VR task in which mice accumulated visual evidence (cues) as they ran down the stem of a T-maze and were rewarded when they turned to the side with more towers at the end. Video frames of the maze are shown below each maze schematic. **(b)** Deep reinforcement learning (RL) network took in video frames from the VR task, processed them with 3 convolution layers, a fully connected layer (orange), and an LSTM layer (green), and outputted an action policy (gray), which inputted the chosen action back into the VR system (blue arrow). The second to last layer of the deep RL network (the LSTM layer) and the weights for the critic served as the inputs to form the Vector RPE (purple). **(c)** Psychometric curve showing the mice's performance (left) and agent's performance (right) after training. The fraction of right choices is plotted as a function of the difference of the right and left towers presented on the trial. For the mice, gray lines denote the average psychometric curves for individual sessions and the black line denotes a logistic fit to the grand mean with bars denoting the s.e.m. (N = 23 sessions). For the model, black bars indicate the s.e.m. **(d)** Deep RL model's scalar value (sum over units in Vector value) during the cue period decreased as trial difficulty (measured by absolute value tower difference, blue gradient) increased.

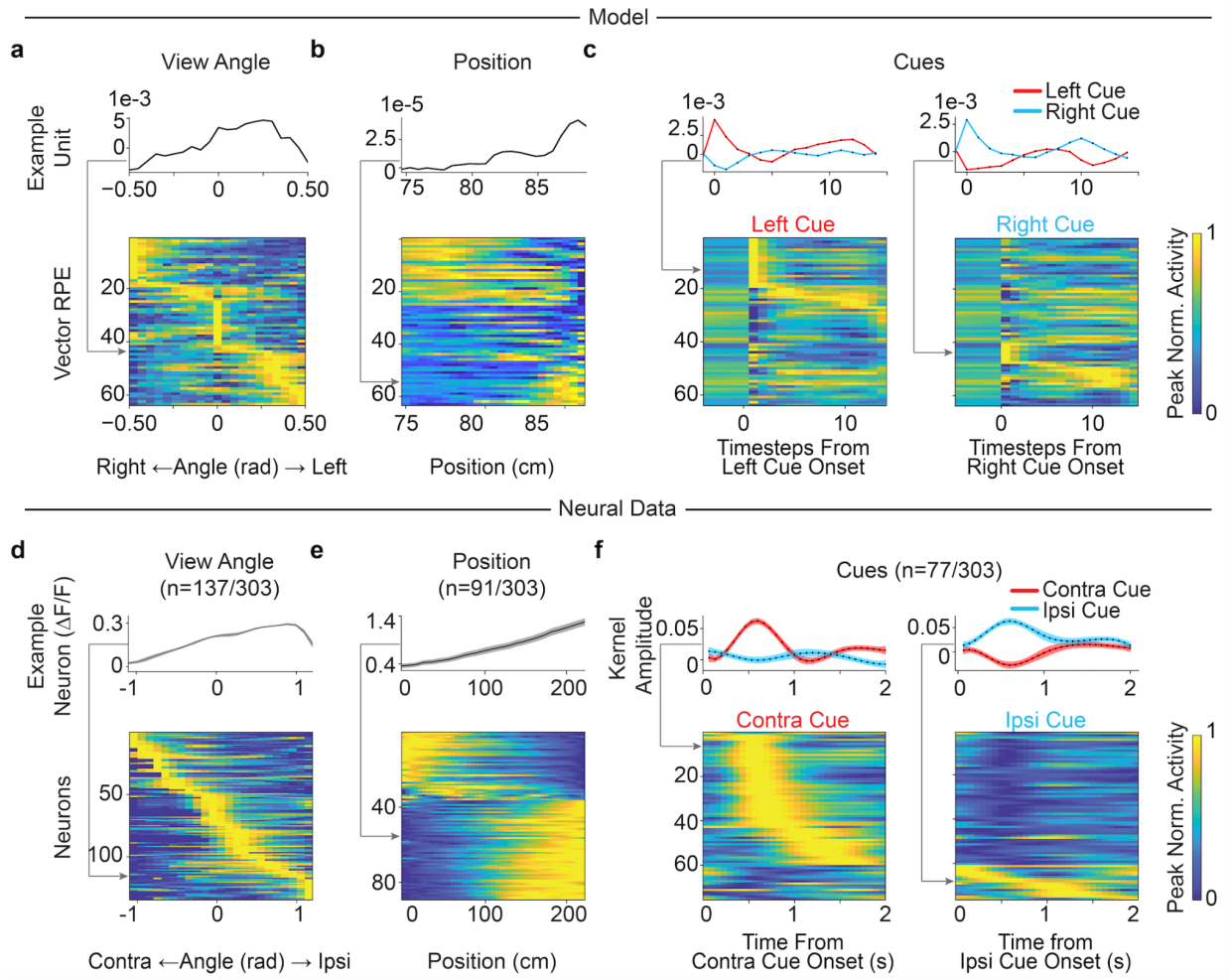


Fig. 3: Vector RPE derived from the trained deep RL model is heterogeneously modulated by behavioral variables during the navigation and decision period, similar to DA neurons. **(a)** Average activity of Vector RPE units plotted with respect to the view angle of the agent. Top panels show an example of a Vector RPE unit's response modulated by each variable and averaged across all trials or cue occurrences; Bottom panels include all Vector RPE units' peak normalized (min-max normalization) activity modulated by the variable, with each row showing a unit's average response and the gray arrow pointing to the example panel's row in the heatmap. **(b-c)** Same as **(a)** but for position of the agent in the final 25 cm of the maze, and left (red) and right (blue) cues. **(d-f)** Same as **(a-c)** but for the subset of neurons from [15] tuned to **(d)** view angle of the mice, **(e)** position, and **(f)** contralateral (red) and ipsilateral (blue) cues. Fringes represent ±1 s.e.m. of averaged signals. In **(d-e)** peak normalized ΔF/F signals are plotted while in (**f**) cue kernels are plotted from an encoding model used to quantify the relationship between the behavioral variables and each neuron.

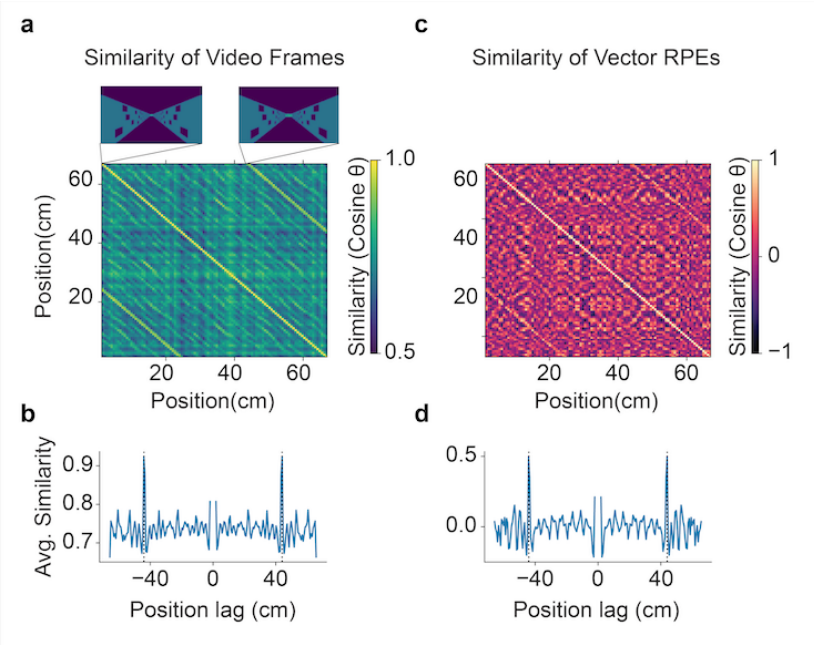


Fig. 4: Vector RPEs reflected incidental high-dimensional visual inputs (a) Similarity matrices of the video frames, which measure the similarity between pairs of video frames (quantified by the cosine of the angle between them when flattened to vectors) across different position combinations. The off-diagonal bands correspond to the wall-pattern repetitions (see video frames for position 0cm and 43cm at insets above). **(b)** Average similarity as a function of distance between frames, indicating that the average similarity peaked at the same position lag (43 cm) for video frames. **(c-d)** same as **(a-b)**, but with vector RPE.

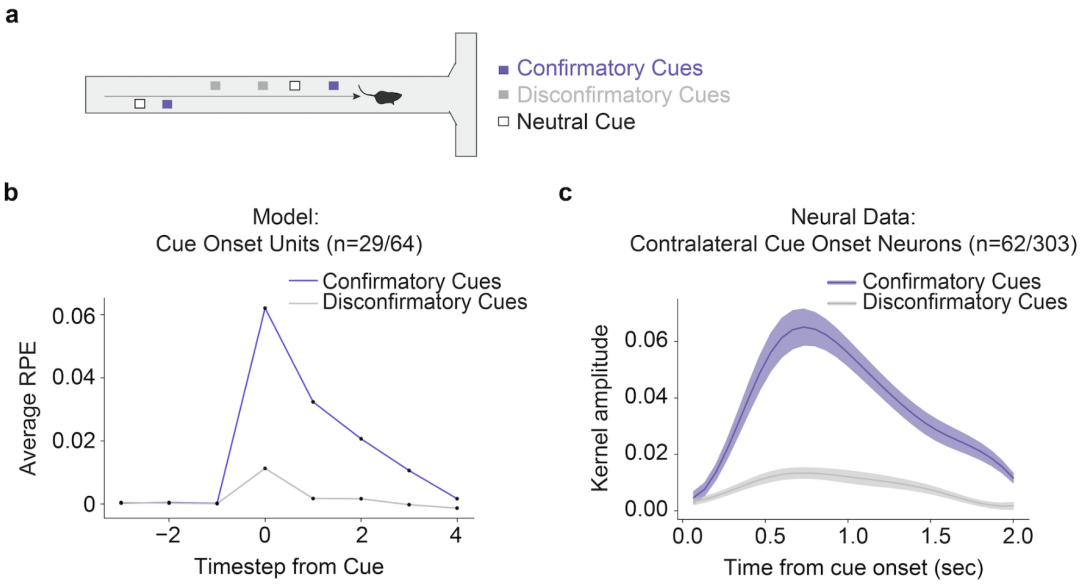


Fig. 5: Cue responses in model and DAergic neurons reflected RPEs with respect to cues, rather than simply their presence. (a) Example trial illustrating confirmatory cues (purple), defined as cues that appear on the side with more evidence shown so far, and disconfirmatory cues (gray), which are cues appearing on the side with less evidence shown so far. Neutral cues (white) occur when there has been the same amount of evidence shown on both sides. **(b)** Average response of to confirmatory (purple) and disconfirmatory cues (gray) for Vector RPE units modulated by cue onset. **(c)** Average responses of the contralateral cue onset DA neurons for confirmatory and disconfirmatory cues. Colored fringes represent ±1 s.e.m. for kernel amplitudes (n = 62 neurons, subset of cue responsive neurons from **Fig. 3f** that were modulated by contralateral cue onset only).

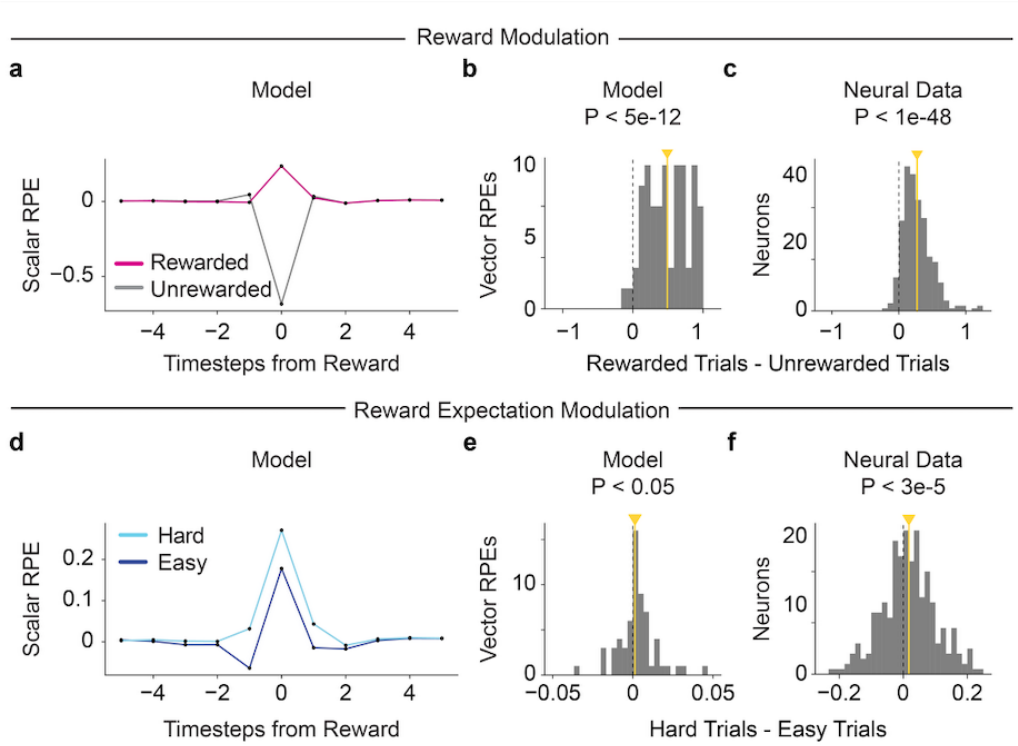
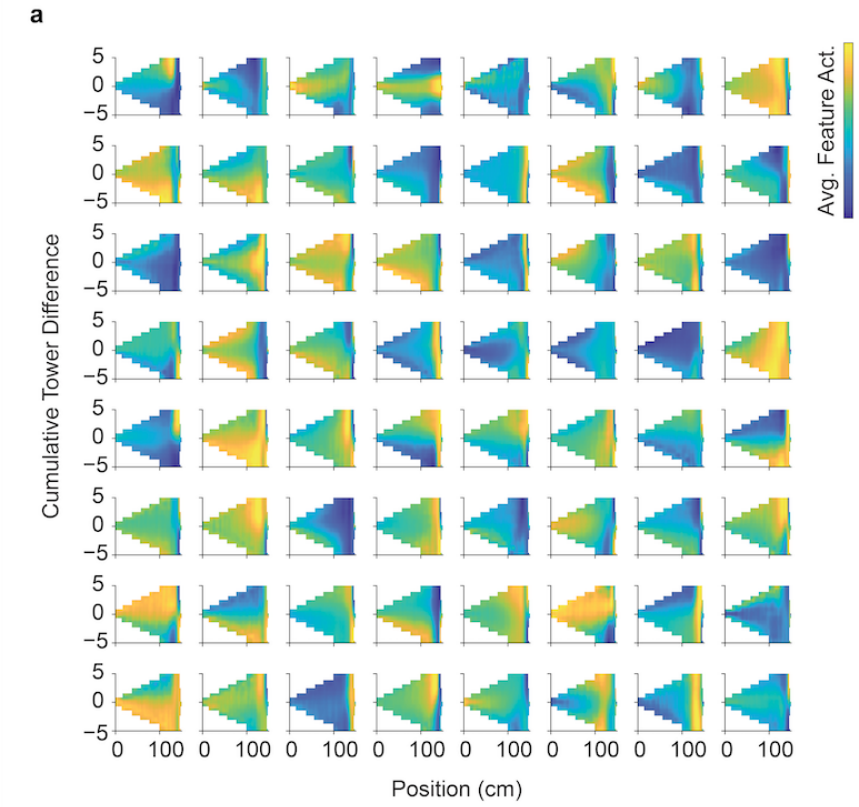
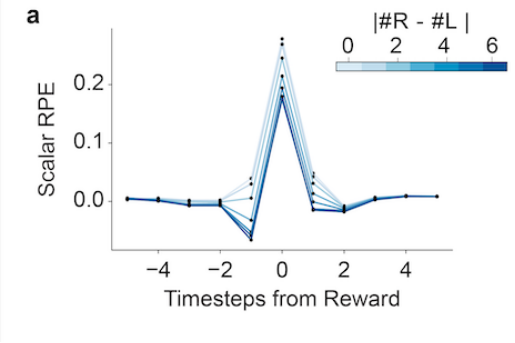


Fig. 6: Vector RPE units and DAergic neurons consistently respond to reward, but show heterogeneous modulation by reward expectation. (a) Average Scalar RPE (sum of units in Vector RPE) time-locked at reward time for rewarded (magenta) minus unrewarded trials (gray). **(b)** Histogram of Vector RPE units' response to reward minus omission at reward time (P < 5e-12 for two sided Wilcoxon signed rank test, N = 64). Yellow line indicates median. **(c)** Same as **(b)**, but with all imaged DA neurons (N = 303), using averaged activity for the first 2 seconds after reward delivery, baseline corrected by subtracting the average activity 1 second before reward delivery (P < 1e-48 for two sided Wilcoxon signed rank test, N = 303). **(d-f)** as in **(a-c)** but for rewarded trials only plotted with respect to trial difficulty (P < 0.05 for two sided Wilcoxon signed rank test, N = 64 in **e** and P < 3e-5 for two sided Wilcoxon signed rank test, N = 303 in **f**). Only reward-responsive neurons are plotted for **f**. Hard (light blue) and easy (dark blue) trials are defined, respectively, as trials in the bottom or top tercile of trial difficulty (measured by the absolute value of the difference between towers presented on either side). In **(e),** there is an outlier datapoint at 0.29 for a Vector RPE unit showing strong reward expectation modulation.



Extended Data Fig. 1: Tuning of 64 LSTM feature units to position and evidence (a) Each panel shows an individual feature unit and how it tunes to the agent's position in the maze and the cumulative tower difference at that position.



Extended Data Fig. 2: Scalar RPE response modulated by the difficulty of the task, defined as the absolute value of the final tower difference (blue gradient) of the trial.

Methods

Behavioral Task

Simulations: At every trial, the agent was placed at the start of a virtual T-maze, with cues randomly appearing on either side of the stem of the T-maze as the agent moved down the maze. On each trial, one side was randomly determined to be correct, and the number of cues on each side was then sampled from a truncated Poisson distribution, with a mean of 2.29 cues on the correct side and 0.69 cues on the incorrect side. In order to prevent the agent from forming a side bias, we used a debiasing algorithm to ensure that the identity of the high probability side changed if the agent kept choosing one side[86]. To match the procedure from the mouse experiment, we also oversampled easy trials (trials in which only one side had 6 total cues) by ensuring that they were 5% of the trials. The agent moved down the maze at a constant speed of 0.638 cm per timestep, and could also modulate its view angle with two discrete actions corresponding to left and right rotation. (A third discrete action moved forward without changing the view angle.) The cue region was 85 cm, and the cues were placed randomly along the cue region with uniform distribution, but with the restriction that cues on either side were constrained to have a minimal spatial distance of 14 cm between them. Each cue first appeared when the agent was 10 cm from the cue location and disappeared once the agent passed the cue by 4cm. After the cue region, there was a short, 5 cm delay region before the agent's final left or right action determined their choice of entering either arm in the T-maze. If the agent turned to the arm on the side where more cues appeared, they received a reward. The agent was also given a sensory input in the model indicating whether it made the correct or wrong turn.

Neural data: The task simulated above was a streamlined version of that used in the mouse recordings in Engelhard 2019. In particular, the rules for spacing and visual appearance of cues were the same, but the simulated controls were simplified (to discrete actions) and the maze was shorter (to facilitate neural network training). Thus the mice could control their speed and direction of movement more continuously by running on a trackball, and in this way traversed a maze that had a 30 cm start region (with no cues), a 220 cm cue region, and an 80cm delay region before the T maze arms. The mean numbers of cues were correspondingly larger: 6.4 on the correct side and 1.3 on the incorrect side. At reward time, the mice received a water reward if they made the correct choice; if a mouse made an incorrect choice, it was given a pulsing 6-12 kHz tone for 1 second. Before the next trial, the virtual reality screen froze for 1 second during reward delivery, and blacked out for 2 seconds if the mouse was rewarded or 5 seconds if the mouse failed.

Virtual Reality System and Deep Reinforcement Learning Model

A deep reinforcement learning network was trained on the evidence accumulation task. As input, the network took in 68 by 120 pixel video frames in grayscale. The model had 3

convolution layers to analyze the visual input, an LSTM layer to allow for memory, and output to 3 action units and 1 value unit. The first convolutional layer had 64 filters, a filter size of 8 pixels, and a stride of 2 pixels; the second convolutional layer had 32 filters, a filter size of 2 pixels, and a stride of 1 pixel; the third convolutional layer had 64 filters, a filter size of 3 pixels, and a stride of 2 pixels. The convolution layers fed into a fully connected layer, which fed into the LSTM layer with 64 units along with a second input, a one-hot vector of length 2 which flagged whether or not the agent was rewarded the end of the trial. The reward input into the LSTM was meant to replicate the sensory input that the mouse experienced when it was rewarded with water or received a tone for failing the trial. The hyperparameters for the convolutional layers were optimized with a grid search of various filter numbers and sizes trained on supervised learning for recognizing towers.

We use the same MATLAB virtual reality program (ViRMEn software engine[87]) from the original neural recordings[15], which we altered to accommodate the agent's movement choices of forward, left, and right. While in the stem of the T-maze, the agent always moved forward at a constant rate per timestep. The constant speed ensured that at every trial, the agent always took the same number of timesteps to traverse the stem of the T-maze. The agent could choose to rotate left or right, which would alter the view angle 0.05 rads up to the limit of -π/6 and π/6 rads. The agent could also choose to move forward without changing their view angle. After the delay region in the T-maze, the agent's left and right movement would no longer alter its view angle, but instead determined which arm the agent chose.

In order for the deep RL agent to interact with the ViRMEn software, we created a custom gym environment using OpenAIGym's *gym* interface[88]. Our custom VR gym environment defined the forward, left, and right movements the agent could make, and sent in the movement choices to the ViRMEn software which in turn returned updated video frames.

We trained the network to maximize obtained reward using the Stable Baselines[89] (version 2.10.1) implementation of the Advantage Actor Critic (A2C) algorithm[41]. All hyperparameters used for the deep RL agent can be found in Table 1. We trained the model until it reached a performance of 80% or higher correct choices, which took 20.8 million timesteps or approximately 130,000 trials.

After training, we froze the weights and took the final output layer of the network before the action and value units (i.e., the LSTM output) as the features for vector value and vector RPE (equation 1, **Fig. 2b**). (Note that this just corresponds to decomposing the scalar value and RPE units in the original A2C network into vectors, with one value component for every LSTM-to-value weight, and a corresponding RPE component: i.e. the Vector RPE model, being algebraically equivalent to TD, is just a more detailed view of the A2C critic). In this way, we calculated the vector RPE at every point in the trial. We defined the outcome period to be the 5 timesteps before and after the reward. We defined the cue period as the first 140 timesteps of the maze, which occurred at the same positions on every trial since the agent always moved forward the same amount at each timestep.

Table 1: Hyperparameters for A2C Algorithm

Parameter (Variable Name in Stable Baselines)	
Number of Parallel Environments (n_env)	8
Number of steps until each environment updates (n_steps)	140
Discount Factor (gamma)	0.99
Learning Rate	0.0025
Value function coefficient for loss calculation (vf_coef)	0.25
Entropy coefficient for loss calculation (ent_coef)	0.01
Maximum value for gradient clipping (max_grad_norm)	0.5
RMSProp decay parameter (alpha)	0.99
RMSProp momentum parameter (momentum)	0.0
RMSProp epsilon (epsilon)	1e-5

Neural Data

This article analyzes data originally reported in Engelhard et al. 2019, the methods of which we briefly summarize here and below[15]. We primarily re-analyzed the the neural recordings during the virtual-reality experiments, in which we used male DAT::cre mice (n=14, Jackson Laboratory strain 006660) and male mice that are the cross of DAT::cre mice and GCaMP6f reporter line Ai148 (n=6 Ai148xDAT::cre Jackson Laboratory strain 030328).

VTA DA neurons were imaged at 30 Hz using a custom built, virtual-reality compatible two photon microscope equipped with pulsed Ti:sapphire laser (Chameleon Vision, Coherent) that was tuned to 920 nm. After imaging, we removed trials in which mice were not engaged in the task, primarily those found close to the end of the session when animal performance typically

decreased. Average performance across sessions on all trials was 77.6 +/- 0.9% after removing trials (compared to 73.3 +/- 1.1% including all trials). Ultimately, we used 23 sessions from 20 mice (one session per imaging field, each session with at least 100 trials and minimal performance of 65%).

After preprocessing the imaging data, we performed motion correction procedures to eliminate spatially uniform motion and spatially non-uniform, slow drifts. The dF/F was derived by subtracting the scaled version of the annulus fluorescence from the raw trace (correction factor of 0.58) and smoothed using a zero-phased filter with 25 point center Gaussian with 1.5 sample points standard deviation. We then divided dF/F by the eighth percentile of the smoothed and neuropil-corrected trace based on the preceding 60s of recording. After examining the dF/F, we only included neurons that were stable for at least 50 trials. The full dataset we used for renalaysis has 303 neurons spread across 23 sessions from 20 mice.

Cue Period Responses

For the heatmaps in **Fig. 3**, each row represents a vector RPE unit or neuron's average response to the behavioral variable rescaled so that their maximum value is at 1 and minimum value is at 0. For the neural data, we used the same encoding model from our previous work[15] to predict neural activity with a linear regression based on predictors such as cues, accuracy, previous reward, position, and kinematics, to isolate temporal kernels that reflect the response to each cue. The code for this encoding model can be found at https://github.com/benengx/encodingmodel. To determine which neurons are displayed for the heatmaps in **Fig. 3d-f,** we used the same criterion as in[15], which was to include neurons with a statistically significant contribution of that behavioral variable in the full encoding model relative to a reduced model, based on an F-test (P = 0.01), with comparison to null distributions produced by randomly shifted data to account for slow drift in the data .

Wall-Texture Analysis

In **Fig. 4**, we identified a repeating wall-texture pattern in the maze by analyzing video frames of a maze without cues with the view angle fixed at 0 degrees. We calculated the similarity matrix for the video frames; specifically, given the video frames, we flattened the video frame at each timepoint into vectors, mean-corrected and normalized the vectors, and measured similarity for all pairs of frames as the the cosine of the angle between these vectors (concretely, the i-jth entry of the similarity matrix gives the cosine of angle between the video frames at times i and j). We also visualized the average, over positions, of each frame's similarity to those ahead and behind it, as a function of distance. We repeated the same analyses on the Vector RPEs, calculating the similarity in the vector RPEs at each timepoint. The Vector RPEs here were derived by running the agent with the trained weights from the normal maze described above on an empty maze without cues, and not allowing the agent to change its view angle in the stem of the maze (always fixed at 0 degrees).

Confirmatory versus disconfirmatory cue responses

We defined confirmatory cues as cues that appeared on the side with more evidence so far, and disconfirmatory cues as cues that appeared on the side with less evidence so far. If the agent or mouse had seen an equal number of cues on both sides, the next cue was defined as a neutral cue. For the neural data, we isolated cue kernels as in Engelhard et al, 2019, with some modifications: instead of using contralateral and ipsilateral cues, we used predictors including contralateral and ipsilateral cues with contralateral evidence, neutral evidence, and ipsilateral evidence so far. For **Fig. 5b,** we selected those vector RPE units that were immediately modulated by cue onset (as opposed to units with a delayed response), regardless of left or right cues. For **Fig. 5c**, we selected among the neurons modulated by cues from the encoding model (N = 77/303), plotting only the units modulated by contralateral cues (N = 62/303).

Outcome Period Responses

In **Fig. 6a,d**, the scalar RPE was calculated by summing the Vector RPE units. For the model responses at outcome time for **Fig. 6b,e**, we took the response at reward time. For the neural responses at outcome time for **Fig. 6c, f**, we matched the original empirical paper[15] and calculated the average activity in the first 2 seconds after the onset of the outcome period, baseline corrected by subtracting the average activity from the 1 second period preceding the outcome. For the histograms in **Fig. 6b-c, e-f**, a two-sided Wilcoxon signed rank test was performed to determine the p value for the median (yellow line).

References

1. Houk, J. C., Adams, J. L. & Barto, A. G. A Model of How the Basal Ganglia Generate and Use Neural Signals that Predict Reinforcement, Models of Information Processing in the Basal Ganglia (eds. JC Houk, JL Davis and DG Beiser), 249/270. (1995).
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. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
4. Reynolds, J. N. J., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).
5. 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).
6. Von Neumann, J. & Morgenstern, O. Theory of games and economic behavior, 2nd rev. (1947).
7. Rescorla, R. A. Learning about qualitatively different outcomes during a blocking procedure. *Anim. Learn. Behav.* **27**, 140–151 (1999).
8. Ganesan, R. & Pearce, J. M. Effect of changing the unconditioned stimulus on appetitive blocking. *J. Exp. Psychol. Anim. Behav. Process.* **14**, 280–291 (1988).
9. Friston, K. J., Tononi, G., Reeke, G. N., Jr, Sporns, O. & Edelman, G. M. Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience* **59**, 229–243 (1994).
10. Matsuda, W. *et al.* Single nigrostriatal dopaminergic neurons form widely spread and highly dense axonal arborizations in the neostriatum. *J. Neurosci.* **29**, 444–453 (2009).
11. Arbuthnott, G. W. & Wickens, J. Space, time and dopamine. *Trends Neurosci.* **30**, 62–69 (2007).

12. Schultz, W. Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27 (1998).
13. 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**, (2019).
14. Choi, J. Y. *et al.* A Comparison of Dopaminergic and Cholinergic Populations Reveals Unique Contributions of VTA Dopamine Neurons to Short-Term Memory. *Cell Reports* vol. 33 108492 (2020).
15. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).
16. Lerner, T. N. *et al.* Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. *Cell* **162**, 635–647 (2015).
17. Collins, A. L. & Saunders, B. T. Heterogeneity in striatal dopamine circuits: Form and function in dynamic reward seeking. *J. Neurosci. Res.* **98**, 1046–1069 (2020).
18. Verharen, J. P. H., Zhu, Y. & Lammel, S. Aversion hot spots in the dopamine system. *Curr. Opin. Neurobiol.* **64**, 46–52 (2020).
19. Hassan, A. & Benarroch, E. E. Heterogeneity of the midbrain dopamine system. *Neurology* vol. 85 1795–1805 (2015).
20. Marinelli, M. & McCutcheon, J. E. Heterogeneity of dopamine neuron activity across traits and states. *Neuroscience* vol. 282 176–197 (2014).
21. Kremer, Y., Flakowski, J., Rohner, C. & Lüscher, C. Context-dependent multiplexing by individual VTA dopamine neurons. *J. Neurosci.* (2020) doi:10.1523/JNEUROSCI.0502-20.2020.
22. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016).
23. Anderegg, A., Poulin, J.-F. & Awatramani, R. Molecular heterogeneity of midbrain

dopaminergic neurons--Moving toward single cell resolution. *FEBS Lett.* **589**, 3714–3726 (2015).

24. Barter, J. W. *et al.* Beyond reward prediction errors: the role of dopamine in movement kinematics. *Front. Integr. Neurosci.* **9**, 39 (2015).
25. Cai, L. X. *et al.* Distinct signals in medial and lateral VTA dopamine neurons modulate fear extinction at different times. *Elife* **9**, (2020).
26. Hamid, A. A., Frank, M. J. & Moore, C. I. Wave-like dopamine dynamics as a mechanism for spatiotemporal credit assignment. *Cell* **184**, 2733–2749.e16 (2021).
27. Wei, W., Mohebi, A. & Berke, J. D. Striatal dopamine pulses follow a temporal discounting spectrum. *bioRxiv* 2021.10.31.466705 (2021) doi:10.1101/2021.10.31.466705.
28. Zolin, A. *et al.* Context-dependent representations of movement in Drosophila dopaminergic reinforcement pathways. *Nat. Neurosci.* **24**, 1555–1566 (2021).
29. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).
30. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**, 85–88 (2012).
31. Haber, S. N., Fudge, J. L. & McFarland, N. R. Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *J. Neurosci.* **20**, 2369–2382 (2000).
32. Alexander, G. E., DeLong, M. R. & Strick, P. L. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annu. Rev. Neurosci.* **9**, 357–381 (1986).
33. Cox, J. & Witten, I. B. Striatal circuits for reward learning and decision-making. *Nat. Rev. Neurosci.* (2019) doi:10.1038/s41583-019-0189-2.
34. Lau, B., Monteiro, T. & Paton, J. J. The many worlds hypothesis of dopamine prediction error: implications of a parallel circuit architecture in the basal ganglia. *Curr. Opin.*

Neurobiol. **46**, 241–247 (2017).

35. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267–279 (2014).
36. Gershman, S. J., Blei, D. M. & Niv, Y. Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
37. Takahashi, Y. K. & Schoenbaum, G. Ventral striatal lesions disrupt dopamine neuron signaling of differences in cue value caused by changes in reward timing but not number. *Behav. Neurosci.* **130**, 593–599 (2016).
38. Takahashi, Y. K. *et al.* Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. *Nat. Neurosci.* **14**, 1590–1597 (2011).
39. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, second edition: An Introduction*. (MIT Press, 2018).
40. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
41. Mnih, V. *et al.* Asynchronous methods for deep reinforcement learning. in *International conference on machine learning* 1928–1937 (jmlr.org, 2016).
42. Voorn, P., Vanderschuren, L. J. M. J., Groenewegen, H. J., Robbins, T. W. & Pennartz, C. M. A. Putting a spin on the dorsal–ventral divide of the striatum. *Trends Neurosci.* **27**, 468–474 (2004).
43. Suri, R. E. & Schultz, W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience* **91**, 871–890 (1999).
44. Gershman, S. J., Pesaran, B. & Daw, N. D. Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *J. Neurosci.* **29**, 13524–13531 (2009).
45. Dabney, W. *et al.* A distributional code for value in dopamine-based reinforcement learning. *Nature* **577**, 671–675 (2020).

46. Matsumoto, H., Tian, J., Uchida, N. & Watabe-Uchida, M. Midbrain dopamine neurons signal aversion in a reward-context-dependent manner. *Elife* **5**, (2016).
47. de Jong, J. W. *et al.* A Neural Circuit Mechanism for Encoding Aversive Stimuli in the Mesolimbic Dopamine System. *Neuron* **101**, 133–151.e7 (2019).
48. 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).
49. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).
50. Lammel, S. *et al.* Input-specific control of reward and aversion in the ventral tegmental area. *Nature* **491**, 212–217 (2012).
51. Syed, E. C. J. *et al.* Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nat. Neurosci.* **19**, 34–36 (2016).
52. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
53. Moss, M. M., Zatka-Haas, P., Harris, K. D., Carandini, M. & Lak, A. Dopamine axons in dorsal striatum encode contralateral visual stimuli and choices. *J. Neurosci.* (2021) doi:10.1523/JNEUROSCI.0490-21.2021.
54. 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).
55. Saunders, B. T., Richard, J. M., Margolis, E. B. & Janak, P. H. Dopamine neurons create Pavlovian conditioned stimuli with circuit-defined motivational properties. *Nat. Neurosci.* **21**, 1072–1083 (2018).
56. Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).
57. Kim, H. R. *et al.* A Unified Framework for Dopamine Signals across Timescales. *Cell* **183**,

1600–1616.e25 (2020).

58. Lammel, S., Lim, B. K. & Malenka, R. C. Reward and aversion in a heterogeneous midbrain dopamine system. *Neuropharmacology* **76 Pt B**, 351–359 (2014).
59. Tsutsui-Kimura, I., Matsumoto, H., Uchida, N. & Watabe-Uchida, M. Distinct temporal difference error signals in dopamine axons in three regions of the striatum in a decision-making task. 2020.08.22.262972 (2020) doi:10.1101/2020.08.22.262972.
60. Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc. Biol. Sci.* **285**, (2018).
61. Kurth-Nelson, Z. & Redish, A. D. Temporal-difference reinforcement learning with distributed representations. *PLoS One* **4**, e7362 (2009).
62. Daw, N. D., Courville, A. C. & Touretzky, D. S. Representation and timing in theories of the dopamine system. *Neural Comput.* **18**, 1637–1677 (2006).
63. Daw, N. D., Courville, A. C. & Touretzky, D. S. Timing and Partial Observability in the Dopamine System. in *Advances in Neural Information Processing Systems 15* (eds. Becker, S., Thrun, S. & Obermayer, K.) 99–106 (MIT Press, 2003).
64. Rao, R. P. N. Decision making under uncertainty: a neural model based on partially observable markov decision processes. *Front. Comput. Neurosci.* **4**, 146 (2010).
65. 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).
66. Lloyd, K. & Dayan, P. Safety out of control: dopamine and defence. *Behav. Brain Funct.* **12**, 15 (2016).
67. Tano, P., Dayan, P. & Pouget, A. A local temporal difference code for distributional reinforcement learning. *Adv. Neural Inf. Process. Syst.* **33**, (2020).
68. Ribas-Fernandes, J. J. F. *et al.* A neural signature of hierarchical reinforcement learning. *Neuron* **71**, 370–379 (2011).

69. Botvinick, M. M., Niv, Y. & Barto, A. G. Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition* **113**, 262–280 (2009).
70. Jiang, L. & Litwin-Kumar, A. Models of heterogeneous dopamine signaling in an insect learning and memory center. *PLoS Comput. Biol.* **17**, e1009205 (2021).
71. Daw, N. D., Kakade, S. & Dayan, P. Opponent interactions between serotonin and dopamine. *Neural Netw.* **15**, 603–616 (2002).
72. Frank, M. J. & Badre, D. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cereb. Cortex* **22**, 509–526 (2012).
73. Estes, W. K. Toward a statistical theory of learning. *Psychol. Rev.* **57**, 94 (1950).
74. Pearce, J. M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
75. Dayan, P. & Long, T. Statistical models of conditioning. *Adv. Neural Inf. Process. Syst.* 117–123 (1998).
76. Rescorla, R. A. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory* 64–99 (1972).
77. Kamin, L. J. Attention-like processes in classical conditioning. in *SYMP. ON AVERSIVE MOTIVATION MIAMI* (1967).
78. Kruschke, J. K. ALCOVE: An exemplar-based connectionist model of category learning. *Psychol. Rev.* **99**, 22–44 (1992).
79. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
80. Gershman, S. J., Norman, K. A. & Niv, Y. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences* **5**, 43–50 (2015).
81. Courville, A. C., Daw, N. D. & Touretzky, D. S. Similarity and discrimination in classical conditioning: A latent variable account. *Adv. Neural Inf. Process. Syst.* **17**, 313–320 (2005).
82. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive

representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* **13**, e1005768 (2017).

83. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
84. Niv, Y. Learning task-state representations. *Nat. Neurosci.* **22**, 1544–1553 (2019).
85. Gustafson, N. J. & Daw, N. D. Grid Cells, Place Cells, and Geodesic Generalization for Spatial Reinforcement Learning. *PLoS Computational Biology* vol. 7 e1002235 (2011).
86. Pinto, L. *et al.* An Accumulation-of-Evidence Task Using Visual Pulses for Mice Navigating in Virtual Reality. *Front. Behav. Neurosci.* **12**, 36 (2018).
87. Aronov, D. & Tank, D. W. Engagement of neural circuits underlying 2D spatial navigation in a rodent virtual reality system. *Neuron* **84**, 442–456 (2014).
88. Brockman, G. *et al.* OpenAI Gym. *arXiv [cs.LG]* (2016).
89. Hill, A. *et al.* Stable Baselines. *GitHub repository* (2018).