What is dopamine doing in model-based reinforcement learning?

Thomas Akam, Mark E. Walton

Department of Experimental Psychology, Oxford University, Oxford, UK

thomas.akam@psy.ox.ac.uk, mark.walton@psy.ox.ac.uk

Abstract:

Experiments have implicated dopamine in model-based reinforcement learning (RL). These findings are unexpected as dopamine is thought to encode a reward prediction error (RPE), which is the key teaching signal in model-free RL. Here we examine two possible accounts for dopamine's involvement in model-based RL: the first that dopamine neurons carry a prediction error used to update a type of predictive state representation called a successor representation, the second that two well established aspects of dopaminergic activity, RPEs and surprise signals, can together explain dopamine's involvement in model-based RL.

Introduction

The reward prediction error theory of dopamine proposes that dopamine neurons signal the difference between how good the future was expected to be, and how good it turned out to be, taking into account both immediate and anticipated long-run rewards[1,2]. As a normative theory that explains a diverse range of experimental data, the influence of RPE theory is hard to overstate. However, an increasing body of findings suggest that RPE is only be a partial account of what dopamine is doing. Signals carried by dopamine neurons appear more heterogeneous than predicted by the theory [3–6], including responses shaped by movement[7–11], and to aversive or threatening stimuli[12,3,13]. A second set of unexpected findings concerns dopamine's involvement in model-based reinforcement learning[14–23] (RL).

Model-based RL uses an internal model of the world that predicts future states given chosen actions, and evaluates the long run value of different options by simulating their likely consequences[2]. This contrasts with model-free reinforcement learning, which stores estimates of the long run value of states and actions, and updates these directly from experience using RPEs. These distinct algorithms have different strengths. Model-based RL uses information efficiently, but is slow and computationally expensive, as it must simulate many possible futures. Model-free RL is quick and computationally cheap, at the cost of less flexible decision making when the environment changes. Brains are thought to use both methods in parallel to exploit the strengths of each[24–26].

As RPE is the key teaching signal in model-free reinforcement learning, RPE dopamine theory was initially conceived with respect to model-free RL. However, several lines of evidence implicate dopamine in model-based RL. First, dopaminergic RPEs do not simply reflect model-free cached values of directly observable events, but rather are informed by inferences based on rich models of the environment[27,14,15,20,22,28]. This is still compatible with an RPE account of dopamine function, but expands the sources of value information that inform RPEs. More troubling for RPE theory is that dopamine neurons respond to events that are surprising but not directly rewarding or aversive, i.e. sensory or state prediction errors (SPEs)[29–32,23,33]. This is fundamentally at odds with RPE theory, because RPEs only occur when outcomes are better or worse than expected. Arguably most strikingly, both causal manipulations[16,18,21,34] and natural variation[17,19] in dopamine function affect model-based learning and decision making. For a detailed review of these findings see Langdon et al [35].

Thus, while there is compelling evidence implicating dopamine in model-based RL, its precise role remains unclear. Here we examine an intriguing recent proposal[36] that dopamine encodes a type of SPE used to learn a predictive state representation called the successor representation (SR). We suggest that some aspects of dopamine anatomy and physiology are hard to reconcile with this account, and propose instead that dopamine's involvement in model-based RL can be explained through two well established aspects of dopamine activity; RPEs and surprise signals.

Dopamine as successor representation prediction error

The basic idea of the successor representation is to learn the long-range predictive relationships between states – i.e. which states generally follow after which other states[37–39]. Specifically, the SR for state X is a vector whose elements indicate the expected discounted future occupancy for each state after starting in state X. This separates the problem of predicting long run value into two components: the expected future states given current state – summarized by the SR – and the immediate rewards available in each state. Long run values can be computed by multiplying the SR with the immediate reward available in each state. This allows value estimates to be updated quickly and simply if the immediate reward available in each state changes (for example due to a change in motivational state or current goal), subject to some important limitations due in part to the dependence of the SR on the specific policy followed while it was being learnt[38]. The SR has attracted substantial recent interest in neuroscience, both as a mechanism for goal-directed behaviour[38,40], and to account for response properties of neurons[41].

As the number of possible discrete states of the world is infinite, practical application of the SR requires working with state features[36,42]. These could be low-level sensory features like colours or high-level abstract features like 'at work'. A feature-based SR is a matrix, mapping the set of features describing the current state of the world onto those features that are likely to be observed in the future. This can be learned using temporal difference methods, based on updating estimates when observations differ from expectations, very similar to those used in model-free RL to learn values. The key difference is that rather than updating a scalar value estimate using a scaler RPE, the SR for a given state feature is a vector indicating how strongly it predicts the future occurrence of each feature (Figure 1A,B). As the prediction is a vector rather than a scalar, the prediction error is also a vector.

The SR account of dopamine activity proposes that dopamine encodes the vector valued SPE used to update a feature-based SR[36]. This accounts for several observations not straightforwardly predicted by RPE theory. Firstly, it predicts heterogeneity of dopamine neuron responses, as different dopamine neurons encode prediction errors for different state features. Secondly, it accounts for dopamine neuron responses to surprising events that are neither appetitive or aversive, as these can still cause SPEs. Thirdly, it predicts that dopamine activity is necessary for learning stimulus-stimulus relationships. Finally, it is argued to explain the ability of dopamine neuron stimulation to drive learning about stimulus-stimulus relationships in situations where it would not normally occur[21,36]. We will return to this last claim later.

Despite these strengths, there are some complications with the SR account of dopamine. The first is dimensionality. The prediction error used to update the SR has the same the dimensionality as the SR itself, i.e. the number of state features. Computational accounts of basal ganglia typically assume that the cortical input represents state, while prediction errors are carried by dopamine neurons. The rat brain contains approximately 17 million cortico-striatal projection neurons[43], but only around 70 thousand midbrain dopamine neurons[44]. Given this vast disparity in neuron numbers, it seems inconceivable that the dimensionality of the dopamine signal is the same as that of the cortical state representation. Furthermore, each dopamine neuron densely innervates a substantial volume of striatum[45], and dopamine neurons communicate predominantly via volume transmission[46,47], further reducing the dimensionality of the dopamine signal received by post synaptic neurons. This massive dimensionality mismatch is not a problem for RPE theory, as the RPE is scalar. It is for an SR account.

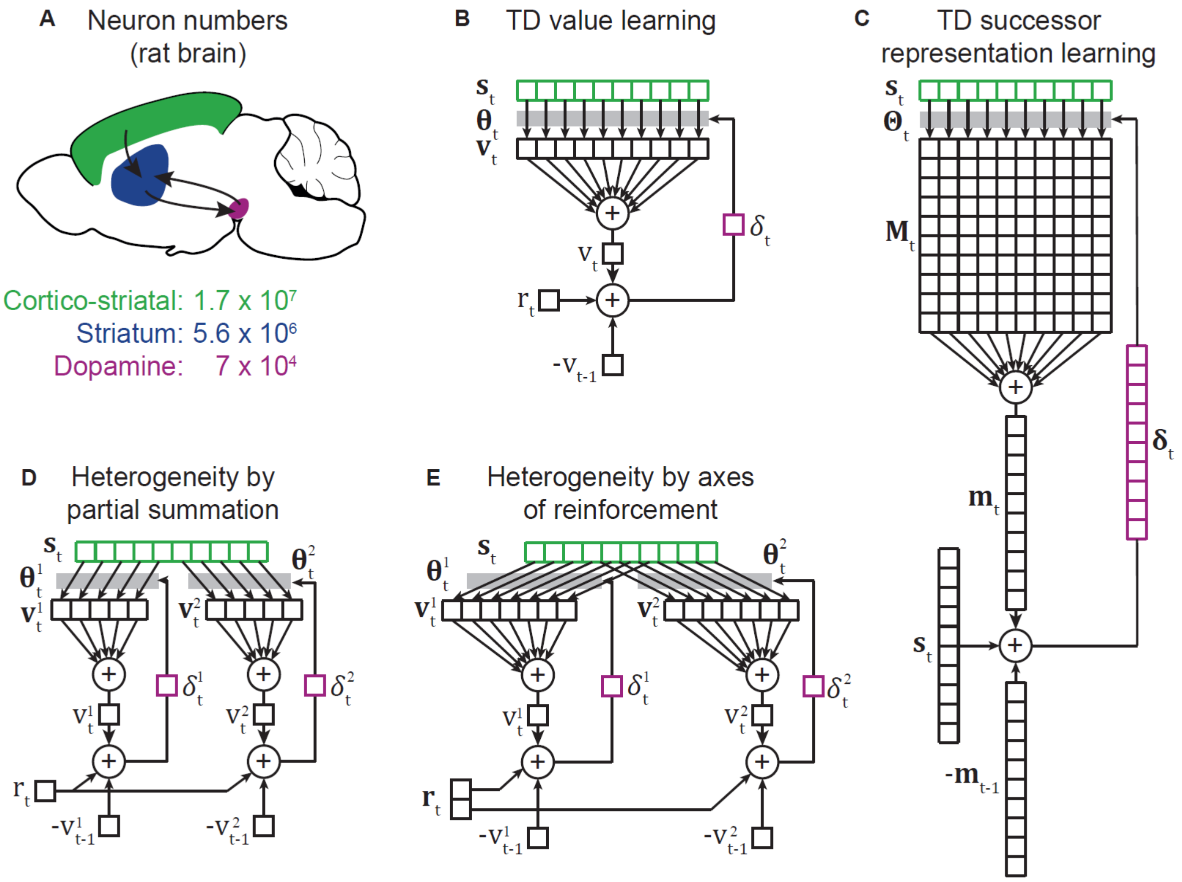

Figure 1. Neuron numbers, signal dimension, and algorithm. A) Number of cortico-striatal, striatal and midbrain dopamine neurons in the rat brain, estimates from [43,44,48]. The disparity in neuron numbers strongly suggests that the signal carried by dopamine is much lower dimensional than the cortical input to striatum. **B)** Diagram illustrating temporal difference (TD) value learning using linear function approximation. A vector $\mathbf{s}_t$ of state features active at time t, is multiplied by a weight vector $\boldsymbol{\theta}_t$ to give a vector $\mathbf{v}_t$**,** whose elements are the contributions made by each state feature to the scalar value v_t of the state. The scalar reward prediction error δ_t, used to update the weights, is computed as $\delta_t = r_t + \gamma v_t - v_{t-1}$, where r_t is the immediate reward at time t, γ is a discount rate and v_{t-1} the value at the previous time step. Irrespective of the dimension of the state representation, the prediction error is a scalar, hence if $\mathbf{s}_t$ is represented by cortical neurons, $\boldsymbol{\theta}_t$by their synapses in striatum, and δ_t by dopamine, the algorithm is consistent with much smaller signal dimension for dopamine relative to cortical input. **C)** Temporal difference learning of a feature based successor representation (SR). The state vector $\mathbf{s}_t$ is multiplied by a weight matrix $\boldsymbol{\Theta}_t$ to give a matrix $\boldsymbol{M}_t$, whose elements $m_t^{i,j}$ are the contributions made by feature i of the current state to predictions about the future occurrence of feature j. Summing the contribution of all features of the current state gives the vector $\boldsymbol{m}_t$, the SR for the current state. As the SR is a prediction of future state features, rather than rewards as in TD value learning, the feature vector $\mathbf{s}_t$ takes the role played by reward r_t in value learning. As the prediction is a vector of dimension given by the number of state features, so is the prediction error $\boldsymbol{\delta}_t$**.** This algorithm does not appear consistent with the massive difference in signal dimension between cortical and dopaminergic input to striatum. **D)** The one dimensional prediction error signal in standard TD value learning is inconsistent with the observed heterogeneity of dopaminergic responses. One possible explanation is that parallel cortico-basal-ganglia loops (labelled 1 & 2) independently learn value estimates, each using only a subset of state features. For clarity we have shown the extreme case of no crosstalk between loops. **E)** Recent data suggest that rather than predicting scalar reward, the basal ganglia predict multiple axes of reinforcement (reward and threat) in loops involving different striatal regions (nucleus accumbens and tail of striatum). Here we have shown two loops which use partially overlapping sets of state features to predict different components of a multi-dimensional reinforcement vector $\mathbf{r}_t$.

Could the dimensionality mismatch be avoided if only a subset of the cortico-striatal input participates in the SR computation? This does not appear to align with the anatomy, as both the cortical and dopaminergic innervation are distributed across the whole striatum. Nor are dimensionality considerations invalidated by emerging evidence for a degree of heterogeneity across dopamine neurons[3–6], or that they carry some information about surprising reward flavours[33] (albeit with weak selectivity) – as the argument is not that the dopamine signal is one dimensional, only that it is much lower dimensional than the cortical input representing state. Accurately estimating the dimensionality of activity is challenging, both because it requires large recorded populations, and because the dimensionality of task related activity is constrained by that of the behavioural task used[49,50]. However, while we lack precise quantitative estimates of the dimensionality of cortical and dopaminergic input to striatum, the neuron number and other considerations outlined above suggest the disparity is large. This constrains the algorithm that the circuit may be implementing.

A second issue is that an SR account does not straightforwardly predict the diverse array of experimental data consistent with RPE theory, including experiments where a large fraction of dopamine neurons qualitatively and quantitatively behaved like an RPE[1,51–54] or that dopamine neuron stimulation can be strongly reinforcing[55–59]. Gardner et al. suggest a resolution to this is to treat reward as just another salient stimulus dimension, such that a subset of dopamine neurons encode a reward prediction error as part of a broader SPE. But the point of the SR is to separate the problem of long run value prediction into one component that learns long run predictions about future states (the SR) and a distinct component that learns about the *immediate* reward available in those states. This necessitates that reward enters into SR based algorithms in a fundamentally different way from state features. Therefore, while RPE-consistent data can be shoehorned into an SR account this way, RPE signals are not predicted by SR theory on normative grounds.

A final issue concerns the proposal that the SR account explains elegant data demonstrating that optogenetic stimulation of dopamine neurons can unblock learning of stimulus-stimulus relationships[21]. In these experiments, rats received presentations of an audio-visual compound stimulus 'AC' that predicted the subsequent presentation of a different auditory stimulus 'X' (denoted as AC→X). If the rats had previously learnt that visual stimulus A alone predicted X (A→X), this blocked learning of the C→X association during subsequent presentation of the compound AC→X. However, if during presentations of AC→X, dopamine neurons were optogenetically activated at the onset of X, learning about C→X occurred. Under an SR account, this could be caused by dopamine stimulation acting as an additional SPE driving learning[36]. The problem with this is the specificity of learning. In order for dopamine stimulation at the time of X presentation to selectively promote learning C→X, the dopamine neurons recruited must be specifically those that represent errors in the prediction of

X. One might argue that these neurons had sub-threshold excitation from the presentation of X itself, and stimulation turned this into super-threshold spiking. However, this seems unlikely as the robust stimulation parameters employed in the experiment would be expected to recruit a large population of dopamine neurons irrespective of any other excitatory input.

Therefore, despite the elegance of the SR account, we think these issues reduce its plausibility, leading us to consider other potential ways that dopamine may interact with model-based RL.

Well informed RPE + surprise = model-based dopamine?

In addition to reward prediction errors, it is well established that dopamine neurons often respond to surprising, novel or salient sensory stimuli[29,31,30,32,23,13,60,61]. We suggest that many observations linking dopamine to model-based RL can be accounted for by the combination of these two aspects of dopamine activity. Specifically, we suggest that dopamine responses to new information comprise an RPE, which takes into account the *best rapidly available* value estimates (which may extend beyond model-free values, as we describe below), and a *scalar*-valued surprise signal. The former can be used to update values stored at cortico-striatal synapses, while the latter is *permissive* for learning state-state predictions but is not in itself a *vector*-valued SPE indicating *how* state predictions should be updated. We will examine each of these ideas in turn.

Dopaminergic RPEs occur at short latency when new information becomes available, so the value information used to compute them must be rapidly available. There are several mechanisms through which internal models could inform value information within the available timeframe: state inference, offline planning, predictive representations, and minimal rollouts (Figure 2).

As current sensory input only partially constrains the current state of the world, brains must infer the world's state by combing recent sensory data with internal models learned over a lifetime. Elegant studies have shown that dopaminergic RPEs reflect inferences about the current state of the world[14,22,28,62], for example inferring reward probability for one stimulus based on sampling another whose reward probability is anticorrelated[14]. Inferences about the world's current state are 'model-based' in the colloquial sense that they depend on internal models. However they are conceptually distinct from 'model-based RL' as defined by Sutton and Barto[2], which refers more narrowly to algorithms which use predictions about future states to compute values[63]. As predictions of future states do appear to shape dopamine responses[15,20], we must consider how model-based RL could inform values within the available timeframe.

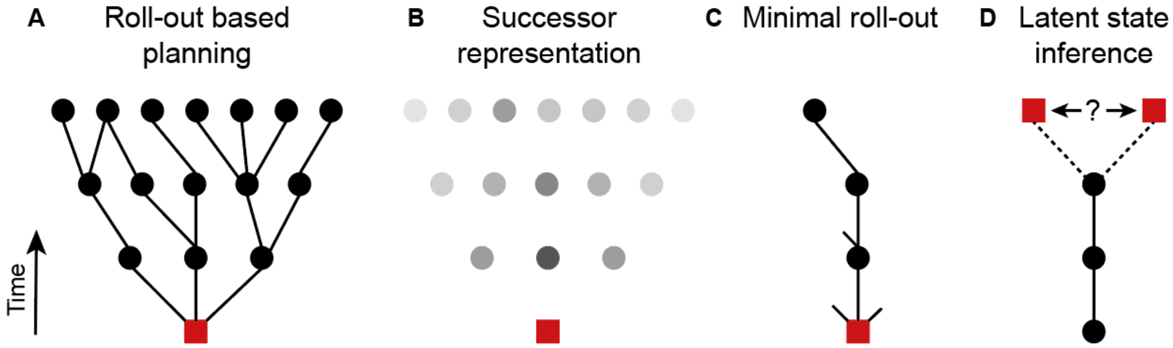

Figure 2. How do internal predictive models contribute to RPE signals? Diagrams showing different ways in which predictive internal models could contribute value information to dopaminergic RPEs. **A)** Model-based planning uses roll-outs along different possible future trajectories (black) from the current state (red) to calculate the long run value associated with different options. The short latency of dopamine responses likely precludes repeated roll-outs using new information from occurring in time to inform the RPE, though offline planning during rest or sleep may update cached values that inform future RPEs. **B)** The successor representation caches a diffuse prediction of likely future states given the current state, which averages over previously experienced behavioural trajectories. This allows for rapid computation of new long run values when the immediate rewards associated with states change, but see [38] for some limitations. **C)** A minimal rollout of only the most probably future trajectory could potentially provide useful value information rapidly in near deterministic settings. **D)** In addition to predicting the future, internal models may help to disambiguate between possible current states that can only be disambiguated by considering the extended history. If the different states have different cached values associated with them, such state inference will affect RPEs.

One possibility is that model-based value computations which contribute to dopaminergic RPEs, do not occur at the time of the RPE, but rather during prior 'offline' planning. This uses simulated experience from a model to refine cached value estimates[2], to improve both future decisions and RPEs that use the cached values. Hippocampal sharp wave ripples, which occur during inactivity and sleep, are thought to implement offline planning[64,65]. During ripples, cell assemblies in hippocampus are activated in sequences representing possible behavioural trajectories[64,66–68]. VTA activity is modulated around ripples, and reward responsive VTA neurons are preferentially recruited when the hippocampal sequence represents reward locations[69]. This ripple associated dopamine activity may update values stored at striatal synapses. A theoretical model which proposes that offline planning is focused on those states that will most improve future choices, explains diverse observations about the content of hippocampal ripples[65]. Behavioural and brain imaging data also suggest that offline planning affects future choices in humans[70,71]. Closed loop electrical stimulation of medial-forebrain bundle (which recruits dopamine neurons) locked to the spiking of a particular place cell during sleep, causes rats to visit the location represented by the place cell on awakening[72], demonstrating artificially induced offline value updating.

The other possibility is that online prediction of the states that will follow an event, allows value associated with them to shape the dopamine response to it. While there is behavioural and neural data consistent with online planning during decision making[73,74], it is unclear whether online planning could inform the dopaminergic RPE to new information. The challenge is that accurate computation of model-based values generally requires repeated roll-outs (i.e. simulated trajectories) along multiple branches of the decision tree[75], which does not appear plausible in time to shape dopamine responses. The successor representation is one possible solution to this problem. It avoids the need for repeated roll-outs by caching a single diffuse prediction of the likely future which averages over trajectories experienced in the past. This allows for rapid value computation by multiplying the SR with the expected immediate rewards in each state. Values estimates derived from an SR could therefore plausibly contribute to dopaminergic RPEs in the available timeframe.

More speculatively, it is worth considering the possibility that a minimal roll-out of the most likely future trajectory could contribute to the dopamine response. Though there is little evidence to suggest sensory events trigger sharp wave ripples at short latency, the hippocampal theta oscillation can be reset by stimuli, particularly when task relevant[76–78]. During each theta cycle, hippocampal neurons activate sequentially, mapping out a trajectory from the current position to that predicted a short time in the future. At decision points, theta sequences explore different possible future paths, suggesting a role in online planning[73,79]. Each theta cycle lasts 120-250ms, while the latency from onset to peak RPE encoding by dopamine is around 150-400ms[80], so theta phase resetting by stimuli could in principle allow the hippocampus to roll-out at least one possible future trajectory informed by the new information. Values derived from a single sample trajectory would be inaccurate when the future was stochastic, but useful in more deterministic settings, particularly if they incorporated an accuracy estimate derived from the predicted trajectories' probability.

Surprise, model-learning, and cortical hierarchy

While there are plausible mechanisms through which model-based value can inform dopaminergic RPEs, this does not explain the observation of dopamine responses to surprising but neutral events[29,31,32,23,61], the necessity of dopamine activity for learning stimulus-stimulus predictions, or that stimulating dopamine neurons can unblock such learning[21]. Consistent with previous accounts[81,82,80], we suggest that dopamine neurons carry a surprise or novelty signal in addition to the RPE, which indicates that something unexpected has happened, independent of valence. We propose that this surprise signal upregulates new learning about predictive relationships, and is responsible for dopamine's causal role in stimulus-stimulus learning.

The anatomy and physiology of the dopamine system suggest a particular arrangement of this surprise signal with respect to cortical hierarchy. Unlike other neuromodulatory systems, dopamine innervation of cortex is most dense to medial frontal and medial temporal lobe regions positioned high in the hierarchy, and comparatively sparse in sensory regions. However, dopamine responses to sensory surprise or novelty can occur at very short latencies (40 – 100ms), comparable to those in V1 and earlier than information about stimulus identity is available in higher cortical regions[82]. This suggests that dopamine neurons carry a surprise signal evaluated low in the sensory hierarchy, close to 'ground truth' sensory data, but transmit it to regions high in the hierarchy working with highly abstracted representations.

Surprise could promote model updating by increasing learning rates and down weighting prior knowledge relative to new information. Consistent with this, dopamine responses to novel conditioned stimuli (CS) promote learning in Pavlovian conditioning, while stimulating dopamine projections to prefrontal cortex during presentation of a familiar CS accelerates learning[61,83]. Additionally, dopamine projections to amygdala mediate surprise induced attention to preceding cues, which promotes subsequent learning about them[84,85]. Upregulation of model learning by dopamine can potentially explain the effect of dopamine manipulations on stimulus-stimulus learning [21,34]. If model updating in frontal regions is upregulated when a prediction failure is signalled by dopamine, inhibiting dopamine neurons will impair stimulus-stimulus learning. Conversely, where prior learning of A→X blocks subsequent learning of C→X during presentation of the compound AC→X, artificially creating a surprise signal at the time of X presentation may down-weight previously learnt predictions, unblocking new learning of C→X. Unlike the SR account of these data, this does not require stimulation to recruit dopamine neurons specific for predictions about X; rather, the surprise signal simply indicates that something unexpected has happened, promoting new learning.

We note that there is substantial overlap between our proposals concerning dopaminergic surprise signals upregulating learning, and previous accounts acetylcholine and noradrenaline function[86]. Exploration of differences between these neuromodulators is beyond the scope of this review. Our aim is rather to point out that surprise signalling in dopamine neurons is well established, that upregulating learning is a normative response to surprise, and hence may account for recent data implicating dopamine in stimulus-stimulus learning.

Mixed signals and heterogeneity

Several questions are raised by our account. Firstly, why mix surprise and RPE signals? One possibility is that though these signals are conceptually very different, the normative responses to each have substantial overlap. Functionally coupling surprise and RPE may drive exploration of undiscovered aspects of the environment, exposing new opportunities[87,88]. Empirically, surprise or novelty can reinforce behaviour, as when rodents press levers to turn on lights[89], a behaviour that is dopamine dependent[90]. Conversely, while any prediction failure suggests that the model should be updated, failure to predict reward has particular behavioural relevance. Coupling RPE and surprise signals may focus the model's representational capacity on those bits of the state-space that are important for behaviour. A second and not mutually exclusive possibility is that there is sufficient separation of these signals in time and/or space to allow downstream regions to demultiplex the two components and differentially respond to them[91].

A second challenge for our account is the increasing evidence for heterogeneity of the dopamine signal, as both RPE and surprise signals are scalar. Computing RPE using a vector valued state representation requires summing the contribution to value estimates from each state feature (Figure 1B). Standard RPE theory assumes that this summation happens prior to the dopamine neurons, so they all carry the same signal. However, given the strong topographical organisation of projections in cortico-basal ganglia loops[92–94], summation prior to dopamine is likely only partial, such that each dopamine neuron combines value information from a subset of state features, generating heterogeneous responses[95] (Figure 1D). As individual dopamine neurons innervate large regions of striatum, a degree of summation will occur downstream of release via volume transmission, though bulk dopamine measurements across different striatal regions indicate that substantial heterogeneity remains at this level[4,13,96].

Though some heterogeneity in a conceptually scalar RPE signal may be accounted for by partial summation of value over state features, it is hard to reconcile the radically different response profile of dopamine neurons projecting to the tail of the striatum, which respond to high intensity sensory stimuli but not unexpected reward, and are aversive when stimulated[13]. Menegas et al. propose that these neurons represent a threat prediction error, forming a separate axis of reinforcement from the value prediction errors carried by dopamine projections to ventral striatum (Figure 1E). Segregating learning about rewards and threats makes sense due to the very different behavioural responses they require. It remains an open question whether other aspects of dopamine heterogeneity, such as movement responses, can be accounted for by partial summation of value information, or indicate the presence of additional axes of reinforcement.

Conclusions

We have proposed that dopamine interacts with model-based RL through three mechanisms. Firstly, dopamine neuron activity during offline planning refines cached values stored at striatal synapses, affecting both future behaviour and RPEs. Secondly, dopaminergic RPEs during behaviour incorporate model-based value information where this is rapidly available; from earlier offline planning, predictive representations such as the SR, and possibly from minimal rollout following stimulus onset. Thirdly, surprise signals carried by dopamine neurons upregulate new learning in predictive models instantiated in cortex and hippocampus. We therefore suggest that, while data implicating dopamine in model-based RL expands the scope of its computational role, and points to a tight integration of model-based and model-free RL mechanisms, it remains compatible with RPE theory and does not require complete re-evaluation of dopamine's role in learning and action.

Acknowledgements:

We thank Neil Burgess and Caswell Barry for information about the effect of stimuli on hippocampal ripples, Marios Panayi for information about lever pressing for lights, and Pete Magill and Paul Dodson for information on numbers of cortico-striatal neurons. Our account of heterogeneity via partial summation over state features borrowed heavily from a talk by Nathaniel Daw at the 2019 Cosyne workshops; any insight there is Nathaniel's, any incoherence our addition. The work was funded by Wellcome Trust awards 214314/Z/18/Z and 202831/Z/16/Z.

References:

1. Schultz, W., Dayan, P. & Montague, P. R. A Neural Substrate of Prediction and Reward. *Science* **275**, 1593–1599 (1997).
2. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction*. (The MIT press, 1998).
3. Lerner, T. N. *et al.* Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. *Cell* **162**, 635–647 (2015).
4. 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).
5. 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).
6. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in VTA dopamine neurons. *Nature* **570**, 509–513 (2019).

*Imaged calcium activity of populations of VTA dopamine neurons while mice performed a decision making task in virtual reality. Dopamine neurons encoded a range of sensory, motor and cognitive variables, with functional clustering of responses in nearby neurons. This heterogeneity is a challenge for RPE theory as RPE is a scalar signal.

7. Syed, E. C. J. *et al.* Action initiation shapes mesolimbic dopamine encoding of future rewards. *Nat. Neurosci.* **19**, 34–36 (2016).
8. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during locomotion and reward. *Nature* **535**, 505–510 (2016).
9. Dodson, P. D. *et al.* Representation of spontaneous movement by dopaminergic neurons is cell-type selective and disrupted in parkinsonism. *Proc. Natl. Acad. Sci.* **113**, E2180–E2188 (2016).
10. 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).
11. 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).

12. Matsumoto, M. & Hikosaka, O. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature* **459**, 837–841 (2009).

13. 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).

** Showed that dopamine neurons projecting to the tail of the striatum respond to novel and high intensity sensory cues but not rewards, and are aversive when stimulated. Proposes that these form a separate 'axis of reinforcement' related to threat prediction.

14. Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values. *J. Neurophysiol.* **104**, 1068–1076 (2010).

15. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).

16. Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine Enhances Model-Based over Model-Free Choice Behavior. *Neuron* **75**, 418–424 (2012).

17. Deserno, L. *et al.* Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proc. Natl. Acad. Sci.* **112**, 1595–1600 (2015).

18. Sharp, M. E., Foerde, K., Daw, N. D. & Shohamy, D. Dopamine selectively remediates 'model-based' reward learning: a computational approach. *Brain* **139**, 355–364 (2016).

19. Doll, B. B., Bath, K. G., Daw, N. D. & Frank, M. J. Variability in Dopamine Genes Dissociates Model-Based and Model-Free Reinforcement Learning. *J. Neurosci.* **36**, 1211–1222 (2016).

20. Sadacca, B. F., Jones, J. L. & Schoenbaum, G. Midbrain dopamine neurons compute inferred and cached value prediction errors in a common framework. *eLife* **5**, e13665 (2016).

21. Sharpe, M. J. *et al.* Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).

22. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589 (2017).

23. Takahashi, Y. K. *et al.* Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron* **95**, 1395-1405.e3 (2017).

24. Balleine, B. W. & Dickinson, A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**, 407–419 (1998).

25. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–11 (2005).

26. Dolan, R. J. & Dayan, P. Goals and Habits in the Brain. *Neuron* **80**, 312–325 (2013).

27. Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y. & Hikosaka, O. Dopamine Neurons Can Represent Context-Dependent Prediction Error. *Neuron* **41**, 269–280 (2004).

28. Babayan, B. M., Uchida, N. & Gershman, S. J. Belief state representation in the dopamine system. *Nat. Commun.* **9**, 1891 (2018).

29. Ljungberg, T., Apicella, P. & Schultz, W. Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.* **67**, 145–163 (1992).

30. Rebec, G. V., Christensen, J. R. C., Guerra, C. & Bardo, M. T. Regional and temporal differences in real-time dopamine efflux in the nucleus accumbens during free-choice novelty. *Brain Res.* **776**, 61–67 (1997).

31. Horvitz, J. C., Stewart, T. & Jacobs, B. L. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Res.* **759**, 251–258 (1997).

32. Lak, A., Stauffer, W. R. & Schultz, W. Dopamine neurons learn relative chosen value from probabilistic rewards. *eLife* **5**, e18044 (2016).

33. Stalnaker, T. A. *et al.* Dopamine neuron ensembles signal the content of sensory prediction errors. *eLife* **8**, e49315 (2019).

34. 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.e3 (2017).

35. Langdon, A. J., Sharpe, M. J., Schoenbaum, G. & Niv, Y. Model-based predictions for dopamine. *Curr. Opin. Neurobiol.* **49**, 1–7 (2018).

* Review detailing recent data implicating dopamine in model-based RL.

36. Gardner, M. P. H., Schoenbaum, G. & Gershman, S. J. Rethinking dopamine as generalized prediction error. *Proc. R. Soc. B Biol. Sci.* **285**, 20181645 (2018).

** Computational study proposing that dopamine neurons encode a type of state prediction error used to update a successor representation. Demonstrates that this idea can account for several experimental observations not predicted by RPE theory.

37. Dayan, P. Improving generalization for temporal difference learning: The successor representation. *Neural Comput.* **5**, 613–624 (1993).

38. 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).

39. Gershman, S. J. The Successor Representation: Its Computational Logic and Neural Substrates. *J. Neurosci.* **38**, 7193–7200 (2018).

40. Momennejad, I. *et al.* The successor representation in human reinforcement learning. *Nat. Hum. Behav.* **1**, 680–692 (2017).

41. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).

42. Barreto, A. *et al.* Successor Features for Transfer in Reinforcement Learning. in *Advances in Neural Information Processing Systems 30* (eds. Guyon, I. et al.) 4055–4065 (Curran Associates, Inc., 2017).

43. Zheng, T. & Wilson, C. J. Corticostriatal combinatorics: the implications of corticostriatal axonal arborizations. *J. Neurophysiol.* **87**, 1007–1017 (2002).

44. Nair-Roberts, R. G. *et al.* Stereological estimates of dopaminergic, GABAergic and glutamatergic neurons in the ventral tegmental area, substantia nigra and retrorubral field in the rat. *Neuroscience* **152**, 1024–1031 (2008).

45. 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).

46. Agnati, L. F., Zoli, M., Strömberg, I. & Fuxe, K. Intercellular communication in the brain: Wiring versus volume transmission. *Neuroscience* **69**, 711–726 (1995).

47. Rice, M. E. & Cragg, S. J. Dopamine spillover after quantal release: Rethinking dopamine transmission in the nigrostriatal pathway. *Brain Res. Rev.* **58**, 303–313 (2008).

48. Oorschot, D. E. Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: A stereological study using the cavalieri and optical disector methods. *J. Comp. Neurol.* **366**, 580–599 (1996).

49. Gao, P. *et al.* A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* 214262 (2017) doi:10.1101/214262.

50. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).

51. Fiorillo, C. D., Tobler, P. N. & Schultz, W. Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science* **299**, 1898–1902 (2003).

52. Bayer, H. M. & Glimcher, P. W. Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron* **47**, 129–141 (2005).

53. Tobler, P. N., Fiorillo, C. D. & Schultz, W. Adaptive Coding of Reward Value by Dopamine Neurons. *Science* **307**, 1642–1645 (2005).

54. Eshel, N., Tian, J., Bukwich, M. & Uchida, N. Dopamine neurons share common response function for reward prediction error. *Nat. Neurosci.* **19**, 479–486 (2016).

55. Tsai, H.-C. *et al.* Phasic Firing in Dopaminergic Neurons Is Sufficient for Behavioral Conditioning. *Science* **324**, 1080–1084 (2009).

56. Witten, I. B. *et al.* Recombinase-Driver Rat Lines: Tools, Techniques, and Optogenetic Application to Dopamine-Mediated Reinforcement. *Neuron* **72**, 721–733 (2011).

57. Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and learning. *Nat. Neurosci.* **16**, 966–973 (2013).

58. Kim, K. M. *et al.* Optogenetic Mimicry of the Transient Activation of Dopamine Neurons by Natural Reward Is Sufficient for Operant Reinforcement. *PLOS ONE* **7**, e33612 (2012).

59. Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**, 117–126 (2016).

60. Kamiński, J. *et al.* Novelty-Sensitive Dopaminergic Neurons in the Human Substantia Nigra Predict Success of Declarative Memory Formation. *Curr. Biol.* **28**, 1333-1343.e4 (2018).

61. Morrens, J., Aydin, Ç., van Rensburg, A. J., Rabell, J. E. & Haesler, S. Cue-evoked dopamine promotes conditioned responding during learning. *Neuron* (2020).

** Showed that dopamine responses to novel stimuli promote learning when they are the CS in Pavlovian conditioning, while stimulating dopamine projections to frontal cortex during presentation of a familiar stimulus CS accelerated learning.

62. Gershman, S. J. & Uchida, N. Believing in dopamine. *Nat. Rev. Neurosci.* **20**, 703–714 (2019).

*Detailed discussion of how uncertainty, including that about the current state of the world, affects reinforcement learning. Reviews evidence that dopamine neurons compute RPEs over inferred states as well as directly observable events.

63. Akam, T., Costa, R. & Dayan, P. Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task. *PLoS Comput Biol* **11**, e1004648 (2015).

64. Buzsáki, G. Hippocampal sharp wave-ripple: A cognitive biomarker for episodic memory and planning. *Hippocampus* **25**, 1073–1188 (2015).

65. Mattar, M. G. & Daw, N. D. Prioritized memory access explains planning and hippocampal replay. *Nat. Neurosci.* **21**, 1609–1617 (2018).

** Developed a normative theory of which state transitions should be replayed during offline planning in order to optimise future choices, and showed that this predicts a diverse set of observations about the context of hippocampal ripples.

66. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006).

67. Diba, K. & Buzsáki, G. Forward and reverse hippocampal place-cell sequences during ripples. *Nat. Neurosci.* **10**, 1241–1242 (2007).

68. Pfeiffer, B. E. & Foster, D. J. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature* **497**, 74–79 (2013).

69. Gomperts, S. N., Kloosterman, F. & Wilson, M. A. VTA neurons coordinate with the hippocampal reactivation of spatial experience. *eLife* **4**, e05360 (2015).

70. Momennejad, I., Otto, A. R., Daw, N. D. & Norman, K. A. Offline replay supports planning in human reinforcement learning. *eLife* **7**, e32548 (2018).

*Used multivariate fMRI analyses to evaluate offline replay of task states in humans. Showed that offline replay is associated with subsequent changes in behaviour, and that surprising outcomes increase subsequent replay.

71. Eldar, E., Lièvre, G., Dayan, P. & Dolan, R. J. The roles of online and offline replay in planning. *bioRxiv* 2020.03.26.009571 (2020) doi:10.1101/2020.03.26.009571.

72. de Lavilléon, G., Lacroix, M. M., Rondi-Reig, L. & Benchenane, K. Explicit memory creation during sleep demonstrates a causal role of place cells in navigation. *Nat. Neurosci.* **18**, 493–495 (2015).

73. Johnson, A. & Redish, A. D. Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *J. Neurosci.* **27**, 12176–12189 (2007).

74. Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D. & Daw, N. D. Model-based choices involve prospective neural activity. *Nat. Neurosci.* **18**, 767–772 (2015).

75. Daw, N. D. & Dayan, P. The algorithmic anatomy of model-based evaluation. *Philos. Trans. R. Soc. B Biol. Sci.* **369**, 20130478 (2014).

76. Givens, B. Stimulus-evoked resetting of the dentate theta rhythm: relation to working memory. *NeuroReport* **8**, 159–163 (1996).

77. Williams, J. M. & Givens, B. Stimulation-induced reset of hippocampal theta in the freely performing rat. *Hippocampus* **13**, 109–116 (2003).

78. Knudsen, E. B. & Wallis, J. D. Closed-Loop Theta Stimulation in the Orbitofrontal Cortex Prevents Reward-Based Learning. *Neuron* **106**, 537-547.e4 (2020).

*Simultaneous recordings in OFC and hippocampus while monkeys learn the values of visual stimuli showed that the hippocampal theta oscillation was reset by trial events and synchronised with OFC. Disrupting the oscillation with microstimulation impaired learning.

79. Kay, K. *et al.* Constant Sub-second Cycling between Representations of Possible Futures in the Hippocampus. *Cell* **180**, 552-567.e25 (2020).

80. Schultz, W. Dopamine reward prediction-error signalling: a two-component response. *Nat. Rev. Neurosci.* **17**, 183–195 (2016).

81. Lisman, J. E. & Grace, A. A. The Hippocampal-VTA Loop: Controlling the Entry of Information into Long-Term Memory. *Neuron* **46**, 703–713 (2005).

82. Redgrave, P. & Gurney, K. The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* **7**, 967–975 (2006).

83. Popescu, A. T., Zhou, M. R. & Poo, M.-M. Phasic dopamine release in the medial prefrontal cortex enhances stimulus discrimination. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3169-3176 (2016).

84. Lee, H. J., Youn, J. M., O, M. J., Gallagher, M. & Holland, P. C. Role of substantia nigra-amygdala connections in surprise-induced enhancement of attention. *J. Neurosci. Off. J. Soc. Neurosci.* **26**, 6077–6081 (2006).

85. Esber, G. R. *et al.* Attention-related Pearce-Kaye-Hall signals in basolateral amygdala require the midbrain dopaminergic system. *Biol. Psychiatry* **72**, 1012–1019 (2012).

86. Yu, A. J. & Dayan, P. Uncertainty, Neuromodulation, and Attention. *Neuron* **46**, 681–692 (2005).

87. Kakade, S. & Dayan, P. Dopamine: generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).

88. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Dopamine modulates novelty seeking behavior during decision making. *Behav. Neurosci.* **128**, 556–566 (2014).

89. Kish, G. B. Learning when the onset of illumination is used as the reinforcing stimulus. *J. Comp. Physiol. Psychol.* **48**, 261 (1955).

90. Olsen, C. M. & Winder, D. G. Operant Sensation Seeking Engages Similar Neural Substrates to Operant Drug Seeking in C57 Mice. *Neuropsychopharmacology* **34**, 1685–1694 (2009).

91. Akam, T. & Kullmann, D. M. Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nat. Rev. Neurosci.* **15**, 111–122 (2014).

92. Watabe-Uchida, M., Zhu, L., Ogawa, S. K., Vamanrao, A. & Uchida, N. Whole-Brain Mapping of Direct Inputs to Midbrain Dopamine Neurons. *Neuron* **74**, 858–873 (2012).

93. Hintiryan, H. *et al.* The mouse cortico-striatal projectome. *Nat. Neurosci.* (2016).

94. Hunnicutt, B. J. *et al.* A comprehensive excitatory input map of the striatum reveals novel functional organization. *eLife* **5**, e19103 (2016).

95. 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).

96. Brown, H. D., McCutcheon, J. E., Cone, J. J., Ragozzino, M. E. & Roitman, M. F. Primary food reward and reward-predictive stimuli evoke different patterns of phasic dopamine signaling throughout the striatum. *Eur. J. Neurosci.* **34**, 1997–2006 (2011).