Reviewed Preprint
v1 • June 22, 2026
Not revised

For correspondence:
luke.priestley@psy.ox.ac.uk
thomas.akam@psy.ox.ac.uk

Competing interests: No competing interests declared

Funding: See page 16

Reviewing editor: Naoshige Uchida, Harvard University, United States

Dopamine ramps as a normative consequence of dual-process control

Luke Priestley[1], Thomas Akam[1,2]

[1]Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom • [2]Sainsbury Wellcome Centre, University College London, London, United Kingdom

eLife Assessment

This **important** study developed a novel theory to account for various aspects of dopamine signals, particularly dopamine ramps. The authors propose that dopamine reward prediction error (RPE) signals are generated by a dual-process learning system in which values inferred by a model-based system enter the RPE asymmetrically into the update target but not the prediction. The results are well-presented and **convincing**, and make a contribution that is of importance to the field. This work will be of interest to those studying dopamine specifically or brain learning computations and systems more broadly.

https://doi.org/10.7554/eLife.111458.1.sa3

Abstract

Midbrain dopamine neurons are thought to implement a temporal difference (TD) reward prediction error (RPE) that updates cached values stored in striatum. This has been challenged by evidence that dopamine "ramps up" to predictable rewards during goal-directed behaviour. Here, we propose that dopamine ramps are RPEs generated by a dual-process learning system in which values inferred using a world model train cached values via the RPE. Ramps arise because efficient training of cached values requires that inferred values contribute to the update target but not the prediction component of the RPE. The model reproduces key dopamine ramp phenomena, including learning dynamics on fast and slow timescales, global updates following changes in reward expectation, transient responses during unexpected state transitions, and sensitivity to state uncertainty manipulations. We therefore argue that dopamine ramps are a signature of interactions between inferred and cached values that revise the traditional dichotomy between model-based and model-free learning.

2 Introduction

Dopamine is widely thought to implement a temporal-difference (TD) reward prediction error (RPE) signal that updates cached values stored at striatal synapses (Montague et al., 1996; Schultz, 2006; Schultz et al., 1997). However, there are features of dopamine activity that this theory struggles to explain. Perhaps most striking is the fact that dopamine in ventral striatum "ramps up" in anticipation of predictable rewards (Howe et al., 2013), particularly in spatial paradigms where distinct locations or sensory states indicate reward proximity (Farrell et al., 2022; Guru et al., 2020; Hamid et al., 2016; Kim et al., 2020; Krausz et al., 2023; Mikhael et al., 2022; Mohebi et al., 2019). The tension with the RPE hypothesis is clear: if dopamine implements an RPE, why do dopamine signals progressively increase during goal approach in well-learned tasks when the value of each state is known?

Theoretical accounts of dopamine ramps fall broadly into two camps. The first proposes that dopamine conveys a value signal rather than an RPE (Hamid et al., 2016; Howe et al., 2013; Mohebi et al., 2019). Although this explains why dopamine might ramp in spatial tasks, where value increases with reward proximity, it is difficult to reconcile with evidence that dopamine exhibits key properties of an RPE in many species and settings (Blanco-Pozo et al., 2024; Eshel et

al., 2015; Kim et al., 2020; O'Doherty et al., 2003; Pessiglione et al., 2006; Schultz et al., 1997; Steinberg et al., 2013; Witten et al., 2011). The second proposes that dopamine ramps are a specific case of RPE that arises under special conditions – for example, when state-uncertainty prevents accurate value estimation (Mikhael et al., 2022), when synaptic decay induces forgetting (Kato and Morita, 2016), or when action-timing is uncertain (Lloyd and Dayan, 2015).

Two, striking, recently reported features of dopamine ramp dynamics are not captured by existing theories. First, new information about expected reward at navigational goals rapidly and globally updates ramp amplitude in a manner that appears inconsistent with TD learning (Guru et al., 2020; Krausz et al., 2023). Second, ramps diminish with experience when animals navigate to the same goal in a stable environment (Guru et al., 2020), but on a much slower timescale than that over which behaviour converges. Guru et al. (2020) and Krausz et al. (2023) propose that model-based value computations underlie the rapid effect of new reward information on dopamine ramps, but a theoretical account explaining why model-based computations give rise to dopamine ramps, and how this explains observed ramp dynamics, is lacking.

Here, we suggest that dopamine ramps are a consequence of a dual-process learning architecture that combines model-based and TD learning systems. This builds on the longstanding idea that the brain has multiple complementary learning systems: an efficient but slow TD system for learning cached values, implemented in the basal ganglia, and a flexible but constrained modelbased system for inferring value using a world-model, putatively implemented in frontal cortex (Balleine and Dickinson, 1998; Daw et al., 2005; Dolan and Dayan, 2013). Our key claim is that if the brain can infer values independently of the TD system, for these to efficiently train cached values they must enter the RPE computation in a specific way that necessarily generates ramps. Specifically: inferred values should contribute to the update target towards which cached values are incremented, but not to the prediction against which outcomes are compared to compute the RPE, which should be determined by cached values alone.

We show that incorporating inferred values into the RPE in this way is normative in the sense that it accelerates learning compared to alternative approaches. We further show that the model generates RPE ramps analogous to dopamine ramps observed in experiments, and reproduces diverse experimental findings on dopamine ramp dynamics. We therefore argue that dopamine ramps are RPEs generated by a normative dual-process learning system, a view that revises the traditional dichotomy between model-free and model-based evaluation.

3 Results

We first review the standard TD learning algorithm, its putative implementation in the brain, and its inconsistency with dopamine ramps. The TD algorithm aims to learn a value function that reflects expected cumulative future reward given a starting state and policy (Sutton and Barto, 2014). Formally:

$$V^{\pi}(s_t) = \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots\right] = \mathbb{E}_{\pi}\sum_{k=0}^{\infty}\gamma^k r_{t+k+1} \quad (1)$$

Where r_t is a reward received at time t, γ is a discount factor, and π is a policy specifying a probabilistic mapping between states of the environment and actions by the agent. For simplicity, we henceforth omit the π superscript from all notation.

If the environment is Markovian, $V(s_t)$ can be expressed recursively as:

$$V(s_t) = \mathbb{E}\left[r_{t+1} + \gamma V(s_{t+1})\right] \quad (2)$$

The TD algorithm exploits this recursion using an online learning rule where value estimates are updated in light of immediate rewards, and differences in value between successive

states. This is formalised in a teaching signal called the reward prediction error δ_t, defined as:

$$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (3)$$

In this equation, the estimate of the old state's value $V(s_t)$ is a prediction – i.e., it is the current 'best guess' about expected future reward – while the immediate reward r_{t+1} plus the discounted value estimate for the new state $V(s_{t+1})$ is an update target – i.e. the value toward which the prediction is adjusted. Value estimates are updated using RPEs as:

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t \quad (4)$$

Where $\alpha \in [0, 1]$ is a learning rate controlling how value estimates change in response to RPEs. The TD algorithm is the basis of an influential account of value-learning in cortico-striatal circuits (Montague et al., 1996; Schultz et al., 1997). This account has three main components fig. 1A-i: (i) The cortex constructs a state representation from sensory experience and communicates it to the striatum (Chang and Tsao, 2017; Liu et al., 2016; Yamins and DiCarlo, 2016); (ii) the striatum estimates value using cortico-striatal synaptic weights, which reflect the relationship between state-features and value (Samejima et al., 2005; Van Der Meer, 2009), and; (iii) VTA dopaminergic neurons compute reward prediction errors that induce plasticity at cortico-striatal synapses (Pawlak and Kerr, 2008), enabling stored values to be updated.

Dopamine ramps challenge this account because the TD algorithm does not, in general, produce RPE ramps in settings where dopamine ramps occur. This is because cached values are assumed to control the agent's policy. In stable and deterministic environments, cached values converge on the true value function with learning, implying that the policy converges when RPEs disappear. This is inconsistent with experimental observations of dopamine ramps, which persist long after animals exhibit expert task performance (Howe et al., 2013; Krausz et al., 2023).

3.1 Inferred values train cached values in a dual process architecture

We propose a dual-process account of dopamine ramps with two core assumptions: (i) that the brain possesses a model-based system that can infer value independently of the basal ganglia TD system in tasks where dopamine ramps occur, and; (ii) that inferred value estimates contribute to training the TD system (fig. 1A-ii). There are many proposals for how the brain might implement model-based value computations (Akam and Walton, 2021; Dolan and Dayan, 2013; Mattar and Lengyel, 2022), including roll-out-based planning, successor or geodesic representations (Dayan, 1993; Sagiv et al., 2025), and inference mechanisms based on attractor dynamics (Donnarumma et al., 2025; Jensen et al., 2025). We do not provide a substantive account of model-based evaluation in this paper. For simulation purposes, we assume a model-based system that infers a goal-conditioned value function using shortest-path distances between states (see below).

If the model-based system can predict future rewards that are not yet reflected in cached values, it can be used to train cached values via the RPE. We argue that to accomplish this, inferred values should contribute to the RPE as:

$$V_{NET}(s_{t+1}) = kV_{MB}(s_{t+1}) + (1-k)V_{TD}(s_{t+1}) \quad (5)$$

$$\delta_t = r_{t+1} + \gamma V_{NET}(s_{t+1}) - V_{TD}(s_t) \quad (6)$$

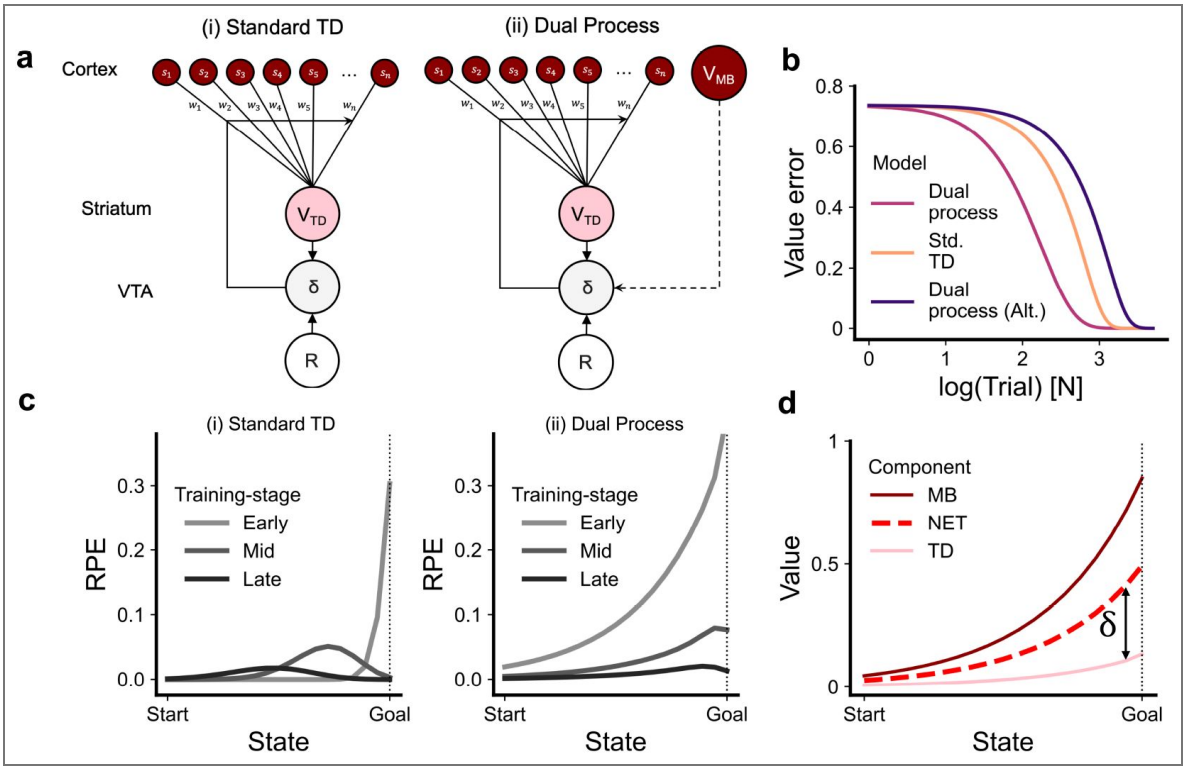

Figure 1. Inferred values train cached values in a dual process architecture.

(a) Diagram of neural circuits for value-learning showing: (i) The standard model of TD learning in cortico-striatal circuits. (ii) The proposed dual-process model. Note the connection from a model-based evaluation system in frontal cortex to dopamine neurons that bypasses striatum. **(b)** Difference between cached values and the true value function (i.e., value error) during learning in the dual-process model, a standard TD agent, and an alternative dual-process model where V_{NET} contributes to both the RPE update target and prediction. **(c)** RPEs during approach to a reward at different stages of learning (early, mid, late) in: (i) a standard TD agent, and; (ii) the dual-process model. **(d)** Value functions in different components of the dual-process model during approach to a reward early in learning. The RPE (δ) is approximately the difference between V_{TD} and V_{NET}.

Here, the estimated value of the new state $V_{NET}(S_{t+1})$ used in the update target combines an inferred value estimate V_{MB} and a cached value estimate V_{TD} using a mixing parameter k. Although our model is agnostic as to how k is determined, it should in principal ensure that V_{NET} reflects the agent's best estimate of the true value given the reliability of V_{TD} and V_{MB} e.g., using uncertainty- or confidence-based arbitration (Daw et al., 2005; Lee et al., 2014). We do not provide a substantive account of arbitration in this paper and our simulations use a fixed value of $k = 0.5$ for simplicity (see Methods). Cached values are then updated using the RPE as:

$$V_{TD}(s) \leftarrow V_{TD}(s) + \alpha_{TD}\delta_t \tag{7}$$

Where α_{TD} is a learning-rate for cached values.

The critical feature of the RPE computation in eq. (6) is that inferred values contribute only to the update target, as $V_{MB}(s_{t+1})$, whereas cached values contribute to both the update target as $V_{TD}(s_{t+1})$, *and* the prediction as $V_{TD}(s)$. The rationale is that because the RPE functions to update cached values, the prediction against which outcomes are compared should reflect cached values alone. However, the update target which cached predictions are updated towards should represent the best available estimate of future reward, and hence should incorporate inferred value estimates if they are available. Computing RPEs in this way updates cached values towards V_{NET}, which is desirable when V_{MB} contains accurate value information that is not yet consolidated into cached values.

How might this dual process architecture be implemented in the brain? A key assumption is that inferred values arise independently of striatal cached values but contribute to RPE computations in the VTA. Since frontal cortex is strongly implicated in model-based evaluation (Akam et al., 2021; Daw et al., 2011; Huang et al., 2020; Jones et al., 2012; Killcross and Coutureau, 2003; Niedringhaus and West, 2022; Stalnaker et al., 2014), we propose that monosynaptic projections from frontal cortex to VTA (Babiczky and Matyas, 2022; Beier et al., 2015; Gao et al., 2022; Wang et al., 2020) communicate the inferred value information used in RPE (fig. 1A-ii, see Discussion).

3.2 Dual process learning generates ramping RPEs

We first evaluated the dual-process model on a 1D tabular environment with a single, absorbing, rewarding goal-state. This mimics the trial-based structure of experiments where dopamine ramps occur, which involve moving through a series of locations to a rewarding goal. Environments with a single, terminal reward yield a special case of the value function where value reduces to the reward available in the goal state discounted by its shortest-path distance from the current state:

$$V_{MB}(s) = \gamma^{d(s)}\hat{r} \tag{8}$$

Where $d(s)$ is the distance between state s and the goal-state and $\hat{r}$ is an estimate of the reward in the goal-state. We assume that distances between states d are known *a priori*. In spatial navigation, this assumption is motivated by entorhinal grid-cells which represent 2D space in a manner that generalises between environments and, in principle, permits distance estimation between locations (Bush et al., 2015; Hafting et al., 2005; Whittington et al., 2020). The estimated reward in the goal-state $\hat{r}$ is updated using a delta rule:

$$\hat{r}_{n+1} \leftarrow \hat{r}_n + \alpha_{MB}\left(r_n - \hat{r}_n\right) \tag{9}$$

Where n indexes a trial of experience, r_n is the reward received on trial n, and α_{MB} is a learningrate for reward estimate updates. We assume that the learning-rate in the model-based system is greater than the learning-rate for the TD system ($\alpha_{MB} > \alpha_{TD}$).

We tested the dual process model's learning performance by comparing time-to-convergence for V_{TD} with respect to the true value function in three different cases: (i) the dual process model proposed in eqs. 5—7, where V_{MB} contributes to the RPE update target but not the prediction; (ii) an alternative dual process model where V_{MB} contributes equally to the RPE update target and prediction, and; (iii) the standard TD learning algorithm, where V_{MB} does not appear (fig. 1B). Cached values converged to the true value function more efficiently in our proposed model compared to both alternatives. Strikingly, the alternative dual-process model, where V_{MB} contributed to both the RPE update target and prediction, learned less efficiently than standard TD. This demonstrates that if inferred values are available, it is normative to use them only in the RPE update target.

We next compared RPEs from the dual-process model and standard TD learning during navigation of the linear track environment (fig. 1C). Ramping RPEs occurred in the dual-process model when inferred values predicted future reward that was not yet captured in cached values. We illustrate this in fig. 1D by displaying V_{TD}, V_{MB}, and V_{NET} during early learning. Inferred values emerged rapidly during initial encounters with the environment, whereas cached values emerged incrementally. Consequently, inferred values exceeded cached values in all states during early learning. Given that V_{MB} contributes only to the update target, and the prediction is given only by V_{TD}, the RPE is shaped by differences between V_{MB} and V_{TD} for successive states, which increase as the agent approaches reward – in other words, RPEs ramp. Dopamine ramps are therefore consistent with inferred values training cached values via the RPE.

3.3 Dopamine ramp dynamics at short and long timescales

A key prediction of the dual process model is that dopamine ramps will diminish over time in stable environments. This is because ramps arise from the difference between cached V_{TD} and inferred V_{MB} values which, in stable environments, (fig. 1) reduces with experience as V_{TD} and V_{MB} converge to the true value function. (fig. 2C). Consequently, RPEs – and therefore ramps – should also reduce with experience. Consistent with this, Guru et al. (2020) report that dopamine ramps in mice gradually diminish with extensive training in a spatial task involving navigation between rewards at alternate ends of a linear track (fig. 2A–B). Dopamine ramps then re-emerged when place-reward contingencies changed, suggesting that ramps implement a learning-related computation consistent with an RPE. RPE ramps in the dual process model reproduced these patterns when simulated on an analogous task (fig. 2D). The dual-process model is thus consistent with long-timescale changes in dopamine ramps reported by Guru et al. (2020).

Guru et al. (2020) further report that dopamine ramps emerge rapidly when rewards are first encountered in a novel environment, (fig. 2E). This is consistent with our model under the assumption that distances between spatial states can be estimated with minimal experience (see above and discussion), and that the model-based system (α_{MB}) employs a high learning-rate for the reward function. Simulating the dual process model in a novel environment confirmed that RPE ramps were absent on the first trial when the the reward at the goal was unknown. Ramps then rapidly developed over subsequent trials as rewards drive learning of inferred values (fig. 2F). The rapid development of dopamine ramps in novel environments is thus consistent with RPE dynamics in the dual-process model.

3.4 Rapid global updates to ramp amplitude by reward

When rewards at goal locations are dynamic, dopamine ramp amplitudes are rapidly updated by reward outcomes. Specifically, Krausz et al. (2023) demonstrate that an outcome at a goal location modulates ramp amplitude on the subsequent visit, with rewards increasing amplitude and omissions decreasing amplitude (fig. 3A–B). Importantly, amplitude changes occur even if the goal is reached by a different route on the subsequent visit, suggesting that they reflect global updates in reward expectation. The dual process model captures these patterns under the assumption that changes in the expected reward at a goal location globally modulate inferred values. This is consistent with inferred values that are computed by combining an estimate of the

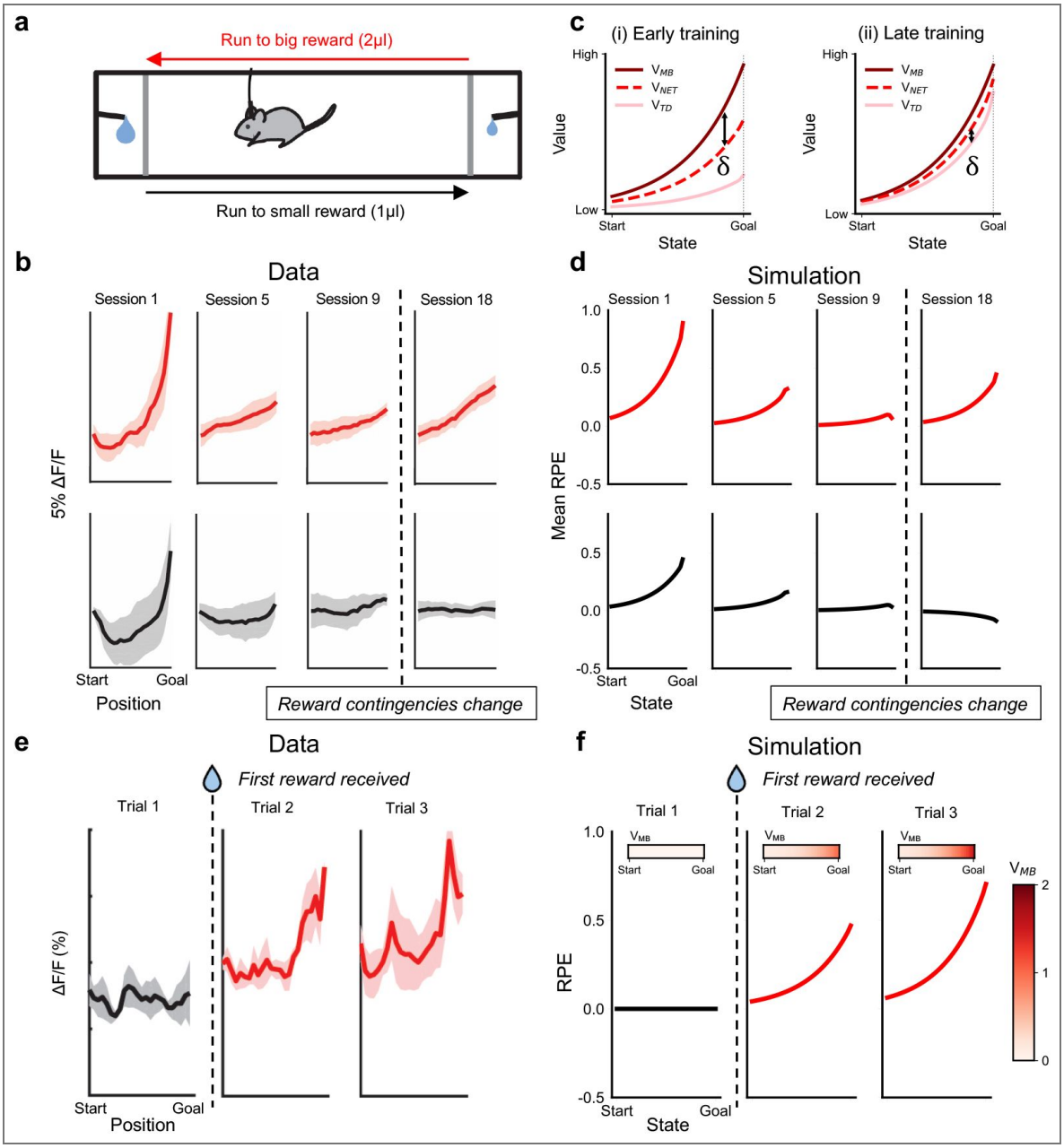

Figure 2. Dopamine ramp dynamics at different timescales.

(a) Diagram of behavioural task in Guru et al. (2020). **(b)** Experimental data from Guru et al. (2020) showing dopamine ramps at different training stages during runs to big (red) and small (black) reward locations. **(c)** Evolution of value estimates during training in the dual-process model; **(i)** value estimates during early training; **(ii)** value estimates during late training **(d)** Evolution of dual process model RPEs during training on a task that replicates Guru et al. (2020). **(e)** Experimental data from Guru et al. (2020) showing rapid development of dopamine ramps after initial encounters with rewarding goals in a novel environment. **(f)** Evolution of RPEs in the dual-process model during initial encounters with rewarding goals.

immediate reward at a goal state with a representation of the distances between states, e.g., via Euclidean distances computed by grid cells (Bush et al., 2015) or cached shortest-path distances as in the Geodesic representation (Sagiv et al., 2025). Implementing the dual process model in a 2D gridworld environment with multiple paths to goal locations recapitulated the patterns reported by Krausz et al. (2023) (fig. 3C–D), suggesting that it is consistent with rapid, global changes in dopamine ramps in environments with dynamic rewards.

3.5 RPE-like dopamine responses to unexpected state transitions

We next tested whether the dual process model reproduces RPE-like dopamine responses during experimental manipulations in spatial tasks. Previous work has shown that when animals progress toward a reward location in a spatial virtual-reality (VR) environment, dopamine signals are modulated by unexpected changes in spatial position (Kim et al., 2020). For example, teleports between non-adjacent states cause dopamine transients that superimpose on dopamine ramps, with magnitudes that are proportional to teleport end-state (fig. 4A) and teleport-distance (fig. 4B). Similarly, the speed at which animals progress towards the goal modulates ramp slopes, with faster speeds producing stronger slopes (fig. 4C). These patterns favour an RPE interpretation of ramps in which dopamine represents changes in value between timepoints, rather than value itself. Dual process model RPEs reproduced dopamine responses during teleport and speed manipulations because RPEs are intrinsically modulated by changes in value (fig. 4A-C). By equating dopamine with an RPE, the dual process model is thus consistent with the results in Kim et al. (2020).

3.6 Dopamine ramps dynamics under state-uncertainty manipulations

Finally, we consider experiments motivated by the proposal that dopamine ramps arise from distortions in learning caused by state uncertainty (Mikhael et al., 2022). As in Kim et al. (2020), mice approached a reward location within a VR corridor (fig. 5A). On some trials, the VR environment progressively darkened during goal approach, thereby increasing the animal's uncertainty about its location (fig. 5A). Dopamine signals on these trials resembled 'bumps' rather than ramps – they initially increased more rapidly than the standard trial signal, before subsequently decreasing below it (fig. 5B).

Following Mikhael et al. (2022), we assume the subject does not know the true current state s_t of the environment, but instead maintains a probability distribution $p(s \mid x_t)$ over possible states s given sensory input x. This probability distribution is assumed to be a Gaussian centred on s_t with standard deviation σ_t:

$$p\left(s \mid x_t\right) \sim \mathcal{N}\left(s_t, \sigma_t\right) \tag{10}$$

Value estimates with respect to x are computed as a probability weighted sum over state value estimates:

$$V_{TD}^{x} \sim \sum_{s} p(s \mid x) V_{TD}^{s} \tag{11}$$

$$V_{MB}^{x} \sim \sum_{s} p(s \mid x) V_{MB}^{s} \tag{12}$$

Simulating this version of the model on standard trials with constant state uncertainty produced ramping RPEs consistent with those observed without state uncertainty (fig. 5C). Following Mikhael et al. (2022), darkening trials were modelled by gradually increasing the width of the state uncertainty kernel across the trial. This generated RPE bumps similar to the dopamine bumps seen in the experimental data (fig. 5C).

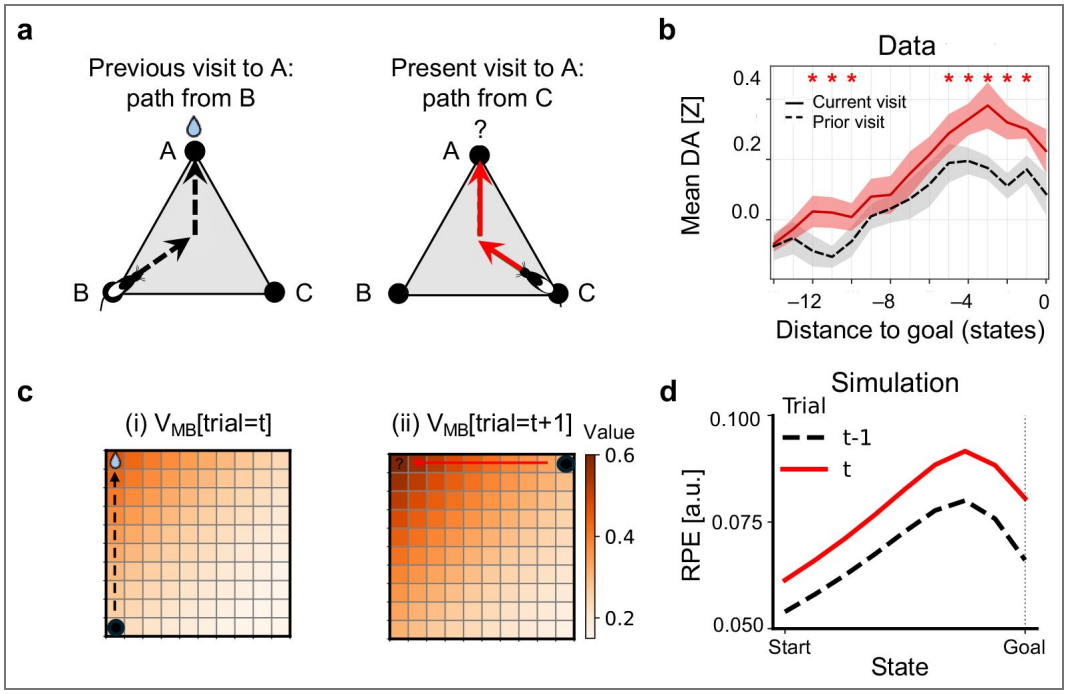

Figure 3. **Reward globally updates dopamine ramps.**

(a) Experimental design from Krausz et al. (2023) in which animals can reach rewarding goal locations via multiple routes. **(b)** Experimental data from Krausz et al. (2023) showing that outcomes at goal locations globally update dopamine ramps regardless of subsequent route. **(c)** Diagram showing global updating of inferred values V_{MB} by rewards in the dual process model. **(d)** Effect of reward on trial *t* on RPEs on trial *t+1* in the dual process model when goal is reached via different routes on *t* and *t+1*.

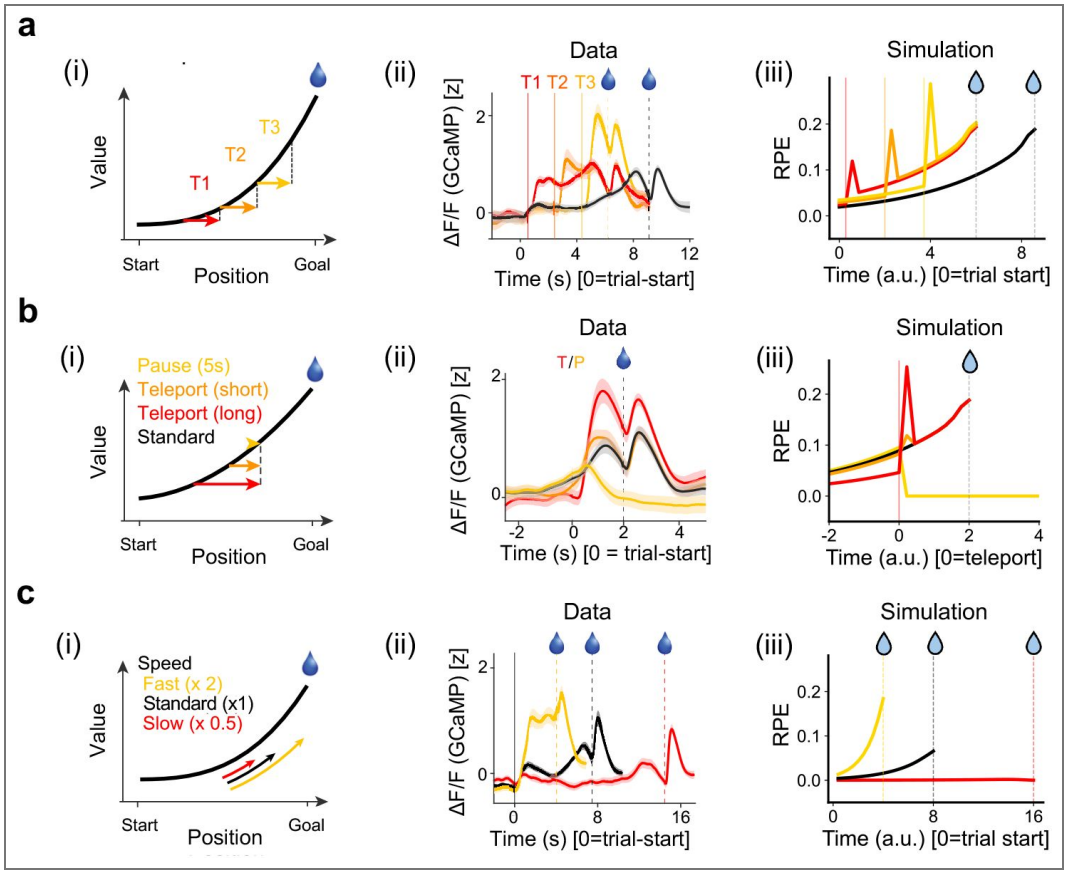

Figure 4. Dopamine responses to unexpected state transitions.

Experimental conditions (left), dopamine recordings (middle), and RPEs in simulations of the dual-process model (right) for key experimental conditions from Kim et al. (2020) in which unexpected state transitions occured during reward approach in a VR environment. **(a)** Teleport end-state manipulation, where teleports of constant distance were aligned with different end-states. **(b)** Teleport distance manipulation, where teleports varied in distance but ended at a common state. **(c)** Traversal speed manipulation.

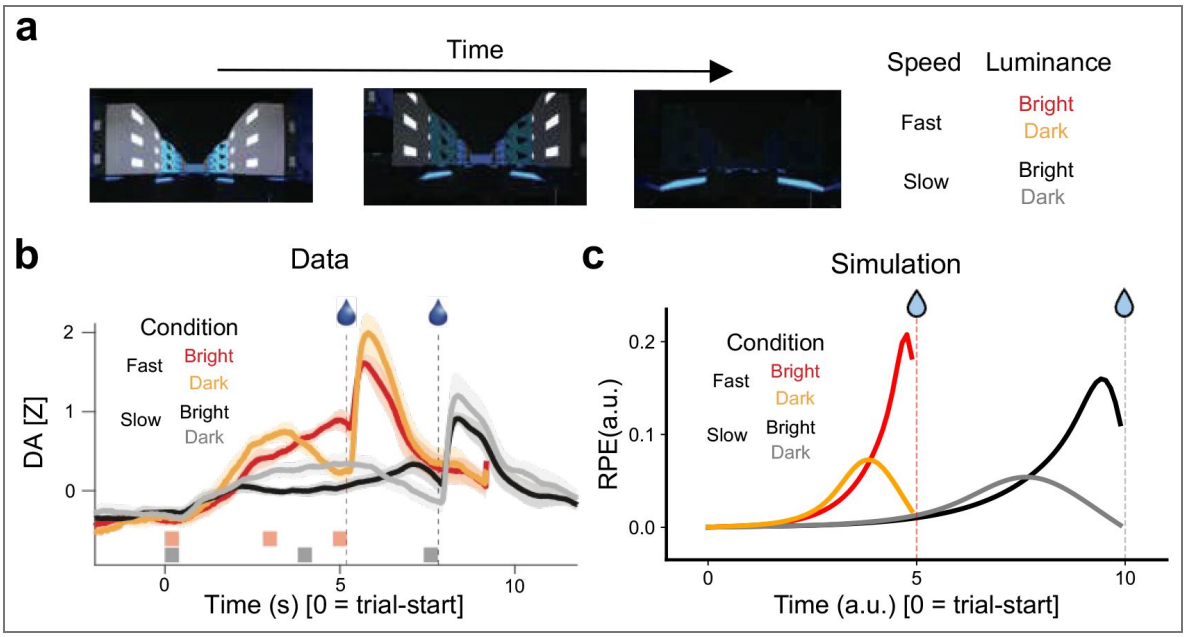

Figure 5. Dopamine ramp dynamics under state-uncertainty manipulations.

(a) Experimental paradigm from Mikhael et al. (2022), where animals approached a rewarded location in a VR environment that differed in movement speed (fast-vs-slow) and luminance (bright-vs-darkening) across trials. **(b)** Dopamine signals from Mikhael et al. (2022), where progressively increasing state uncertainty causes dopamine bumps rather than dopamine ramps. **(c)** Effect of progressively increasing state uncertainty on RPEs in the dual process model.

RPE bumps occur because state uncertainty distorts the inferred value estimates V^x_{MB} that mdrive RPEs. Greater uncertainty assigns more weight to states away from the true state. Early in the trial, this increases inferred value estimates, as although the uncertainty kernel is symmetric around the true state, the slope of the value function is steeper on the higher value side. Late in the trial, uncertainty adds probability mass primarily behind the true state, as the uncertainty kernel cannot extend beyond the reward location if the reward has not been reached. This decreases inferred value estimates and hence the RPE.

Although state uncertainty manipulations have similar effects in both our model and Mikhael et al. (2022), the underlying mechanism of RPE ramps is fundamentally different. Mikhael et al. (2022) propose that RPE ramps arise from a correction term in the value update which counteracts biases that arise from state uncertainty. Learning therefore converges not when the RPE is zero but rather when it is cancelled out by the correction term. Since the required correction is proportional to state value, this generates RPE ramps. This account rests on the critical assumption that state uncertainty is systematically larger when estimating the value of the new state V_{xt+1} compared to the value of the old state $V_{x\,t}$ in the RPE computation – the rationale being that sensory feedback reduces uncertainty for the old state s_t relative to the new state s_{t+1}. It is this assumption that causes state uncertainty to systematically bias learning, and necessitates the correction term that generates RPE ramps. Importantly, recent experimental data measuring the effect of striatal stimulation on dopamine signals (Campbell et al., 2025) suggests that temporal difference value comparison is implemented by synaptic delays in striatum-VTA circuitry, such that $V_{x\,t}$ is simply a delayed copy of $V_{x\,t+1}$, and hence inherits the same state uncertainty.

In contrast to Mikhael et al. (2022), our model reproduces dopamine dynamics under uncertainty manipulations without requiring systematic differences in state uncertainty between terms in the RPE computation.

4 Discussion

Dopamine ramps have attracted widespread interest because they appear to contradict the theory that dopamine implements a temporal difference reward prediction error (Berke, 2018; Hamid et al., 2016; Howe et al., 2013; Niv, 2013). Here, we propose that dopamine ramps are RPEs generated by a dual-process learning architecture in which values inferred from a world model train cached values via the update target of the RPE. We show that this architecture accelerates cached value learning, and reproduces the key empirical features of dopamine ramps within a unified explanatory framework.

The dual process model generates ramping RPEs because inferred values contribute to the update target but not to the prediction component of the RPE. The rationale is that the update target should represent the best estimate of future reward, and inferred values should therefore contribute to it if they are accurate. The prediction, by contrast, should be determined by cached values alone, because cached values are the quantities that the RPE must update. This asymmetric use of inferred values is normative in the sense that, given accurate inferred values, it accelerates convergence of cached values to the true value function. Strikingly, incorporating inferred values symmetrically in both the update target and prediction slows learning relative to not using inferred values at all (fig. 1C).

The model accounts for key experimental findings on dopamine ramp dynamics via the interplay of two sources of value information (figs. 2 to 4): (i) fast-evolving inferred values, which explain rapid global updates following individual rewards, and (ii) slow-evolving cached values, which explain why ramps diminish with experience as cached values incrementally converge. This clarifies why ramps persist after expert behaviour has developed, as policy can be guided by inferred values long before cached values converge.

Our theory implies that the brain has mechanisms for inferring value online during behaviour, independently of the striatum. This contrasts with offline replay mechanisms that refine cached value estimates guiding subsequent behaviour (Mattar and Daw, 2018; Sutton, 1991). Since

modelbased evaluation is generally computationally intensive (Sutton and Barto, 2014), this raises the question of how online inferred value estimation is tractable. We suggest that fast, online inferred value estimation is possible only in specific situations that permit efficient solution methods. Specifically, RL problems characterised by absorbing goal-states reduce goal-conditioned value functions to the immediate reward at the goal, discounted by the distance or cost to reach it. This enables value to be estimated using learned distances between locations (Piray and Daw, 2021; Sagiv et al., 2025). Although real-world behaviour continues after goals are reached, the brain might use the solution methods afforded by absorbing goal states as a heuristic, or as a component of a hierarchical control architecture (Ringstrom et al., 2025).

These considerations suggest that dopamine ramps will emerge when: i) the brain has an internal model of distances between states, and ii) behaviour is organised by discrete, known, and rewarding goal states, rather than random foraging. Goal-directed spatial navigation exemplifies these conditions, explaining the prominence of ramps in navigation tasks. Grid cells facilitate distance estimation in physical space and carry representations that generalize across environments (Bush et al., 2015; Hafting et al., 2005; Whittington et al., 2020), consistent with the rapid onset of dopamine ramps in spatial tasks (Guru et al., 2020). Hippocampal-entorhinal circuits for spatial cognition also represent position in sensory or abstract state spaces (Aronov et al., 2017; Constantinescu et al., 2016), which may explain ramps in tasks where sensory cues indicate reward proximity (Kim et al., 2020). By contrast, classical conditioning tasks lack distinct sensory states indicating reward proximity. Dopamine responses in these settings resemble a backpropagating TD error rather than a ramp, consistent with inferred values playing no role in the update target (Cohen et al., 2012; Schultz et al., 1997).

Neural implementation of the dual process architecture requires that dopamine neurons receive inferred value information through a non-striatal pathway. Although the neural basis of model-based evaluation remains poorly understood, it has been linked to orbitofrontal (OFC) and medial frontal cortex (mFC) in both humans and non-human animals (Akam et al., 2021; Daw et al., 2011; Huang et al., 2020; Jones et al., 2012; Killcross and Coutureau, 2003; Niedringhaus and West, 2022; Stalnaker et al., 2014). Notably, mFC and OFC have monosynaptic projections to VTA dopamine neurons (Babiczky and Matyas, 2022; Beier et al., 2015; Gao et al., 2022; Wang et al., 2020), and stimulating the FC-VTA pathway induces conditioned place preference via the nucleus accumbens (Beier et al., 2015). VTA-projecting FC neurons are therefore a plausible source of the inferred value information that generates dopamine ramps (Guru et al., 2020). Our model further implies that VTA-projecting and striatum-projecting subpopulations of FC neurons should encode different signals: inferred values in the former, and state features in the latter. Consistent with this, these subpopulations are largely anatomically separate, although their coding properties remain uncharacterised (Babiczky and Matyas, 2022; Gao et al., 2022).

Our model makes several testable experimental predictions. (1) VTA-projecting frontal cortex neurons will encode inferred value signals (i.e., expected discounted future reward) in settings where dopamine ramps occur. (2) Silencing the FC-to-VTA pathway, or the components of the world model necessary for inferred value estimation, will abolish dopamine ramps. (3) Abolishing dopamine ramps will slow the development of striatal state-value signals during goal-directed navigation (Van Der Meer, 2009), and alter the development of state value representations during learning. (4) Transiently stimulating FC-to-VTA and striatum-to-VTA pathways will evoke distinct patterns of ventral striatal dopamine release. Stimulating NAc D1 neurons initially excites then subsequently inhibits VTA neurons (Campbell et al., 2025), consistent with the dual role of cached values in the RPE update target and prediction. Stimulating the FC-VTA pathway should, by contrast, evoke VTA excitation alone, since inferred values contribute only to the update target.

Finally, together with Mattar and Daw (2018), our work suggests that the traditional dichotomy between model-based and model-free evaluation should be revised. Each account emphasises that model-based mechanisms have a profound influence on the striatal cached value system – online through dopamine ramps, and offline through replay. This implies that the cached value system should not be viewed as model-free, but rather as a long-term memory system for value that is shaped by both temporal-difference learning and model-based evaluation.

6 Methods

6.1 General simulation details

All simulations were implemented in Python v3.14. For simplicity, the policy for all agents in all simulations was deterministic and involved moving directly to the rewarding goal location. Dualprocess agents were simulated according to eqs. 5—7. Task specific environments and parameter choices are described below.

Code for replicating the simulations and generating the manuscript figures is available at: *https://github.com/lpriestley/da-ramps*

6.2 Comparison of value-learning algorithms

To characterise whether the dual-process model accelerated value-learning (fig. 1B), we implemented (i) a dual-process agent, (ii) a standard TD agent, and (iii) an alternative dual-process agent on a linear track environment. The standard TD agents was implemented according to eqs. 3—4. The alternative dual process agent was implemented according to:

$$V_{NET}(s) = kV_{TD}(s) + (1-k)V_{MB}(s) \tag{13}$$

$$\delta_t = r_{t+1} + \gamma V_{NET}(s_{t+1}) - V_{NET}(s_t) \tag{14}$$

$$V_{TD}(s) \leftarrow V_{TD}(s) + \alpha_{TD}\delta_t \tag{15}$$

The key difference compared to the dual process agent defined in eqs. 6—7, therefore, is that inferred values appear in both the RPE update target and the prediction, instead of the update target alone. This alternative dual-process agent learned slower than standard TD learning (fig. 1B) and did not generate ramping RPEs. Parameters for the agents were: α_{TD} = 0.01, α_{MB} = 0.50, γ = 0.93, k = 0.50,

The linear track was formalised as a tabular environment with N = 10 states. There was a goal state at one end of the track with a scalar reward r = 1.0. Agents started each trial at the end of the track opposite the goal state. All agents performed the task for T = 5000 trials. Value error was calculated on each trial as $\frac{1}{N}\sum_{s=1}^{N} V(s) - V_{TD}(s)$ — i.e the average discrepancy between cached values V_{TD} and the true value function V.

6.3 Characterising dual-process learning dynamics

To demonstrate why the dual-process model produces ramping RPEs, a dual-process agent was simulated on a linear track, formalised as a 1D tabular environment with N = 20 states, and a goal state at one end of the track with a scalar reward r = 1.0. Agents started each trial at the end of the track opposite the goal state. The parameters for the agent were: α_{TD} = 0.01, α_{MB} = 0.50, γ = 0.85, k = 0.50. The value-functions V_{MB}, V_{TD} and V_{NET} were extracted from the agent on trial t = 100 of learning and displayed in fig. 1D. RPEs δ at early, intermediate and late stages of learning were further extracted and graphed in fig. 1E-ii. This was compared to RPEs from a standard TD agent simulated with parameters α_{TD} = 0.01, γ = 0.85.

6.4 Guru et al. (2020) simulation

We compared dual-process RPEs to dopamine signals in Guru et al. (2020). The Guru et al. (2020) study involved recording dopamine signals whilst mice ran between alternate ends of a linear track. One end of the track had a large reward (2μL), and the other end of the track had a small reward (1μL).

To simulate the dual-process model on this task, we treated navigation towards each end of the track as a separate state space, consistent with hippocampal units having strong movement direction tuning on linear tracks. Each agent was simulated with the following parameters: α_{TD} = 0.005, α_{MB} = 0.50, γ = 0.93, k = 0.50. Each linear track was formalised as a 1D tabular environment with N = 35 states and a goal state at one end. Agents started each trial at the end of the track opposite the goal state. In the large-reward track, the initial reward value was r = 2.0, and the small-reward track, the initial reward value was r = 1.0.

Agents performed S = 18 sessions of learning, where each session involved T = 100 trials. On session s = 17, the reward values at the end of high-reward and low-reward tracks were swapped. The training regime was designed to replicate the Guru et al. (2020) experiment.

In fig. 2C, V_{MB}, V_{TD} and V_{NET} were extracted from trial t = 100 on sessions $s \in \{1, 4\}$. In fig. 2D, the evolution of RPEs over learning was visualised by computing, for each state, the mean RPE over trials within a session. In fig. 2F, RPEs and V_{MB} were extracted for trials $t \in \{1, 2, 3\}$.

6.5 mKrausz et al. (2023) simulation

We compared dual-process RPEs to dopamine signals in Krausz et al. (2023). In Krausz et al. (2023), rats performed a maze navigation task, in which a series of goal locations delivered probabilistic rewards. The task took place in a complex maze environment with multiple pathways to each goal location.

To simulate the task, a dual-process agent was implemented on a a 2D 10 × 10 gridworld. There was a goal-location g = (1, 1) which, when visited, delivered reward stochastically with r = 1.0 and $p(reward)$ = 0.5. The agent was alternately started on odd and even trials from s_{odd} = (10, 1) and s_{even} = (1, 10) and followed a trajectory directly to the reward location. This allowed us to test how specific rewards and omissions at the goal location influenced RPEs on the subsequent trial, even when the trajectory to the goal location was different. The agent was simulated with the following parameters: α_{TD} = 0.01, α_{MB} = 0.10, γ = 0.85, k = 0.50. It performed T = 500 trials. In fig. 3C, the effect of rewards on inferred values was visualised by extracting V_{MB} for consecutive trials $t - 1$ and t, where $t - 1$ was rewarded. In fig. 3D, the effect of rewards on RPEs was visualised by comparing RPEs on consecutive trials $t - 1$ and t, where $t - 1$ was rewarded.

6.6 Kim et al. (2020) simulation

We compared dual-process RPEs with dopamine signals in Kim et al. (2020). In the Kim et al. (2020) experiment, subjects viewed a VR track environment with a terminal reward at the end of the track. The environment was manipulated using teleports between non-adjacent states, and speed modulations that controlled how quickly subjects moved through the environment.

We simulated the dual-process agent on these experiments using linear tracks, which were formalised as 1D tabular environments with a goal-state that delivered a reward r = 1.0 at one end of the track. The agent always started at the end opposite the goal. In the teleport experiments (fig. 4A and fig. 4B), the track had N = 32 states. In the speed-manipulation experiment (fig. 4C), the track had N = 40 states. In the teleport-distance experiment, all teleports ended at state $s_{teleport\text{-}destination}$ = 24, where short teleports had distance d_{short} = 2 and long teleports had distance d_{long} = 10. In the pause condition, the agent remained in $s_{teleport\text{-}destination}$ for an arbitrary number of timepoints. We abolished the effect of inferred values on RPEs during the pause period under the assumption that inferred values predict temporally discounted future reward. We assume that such predictions are null in situations when the agent is static in a non-rewarding state. In the teleport end-state experiment, teleports had a constant distance d = 10 and were initiated from either early, moderate, or late start locations where s_{early} = 2, $s_{intermediate}$ = 10, s_{late} = 14. In the speed-manipulation experiment, slow, normal and fast speeds were implemented by modulating the step-size with which the agent moved through the environment, where $stepsize_{small}$ = 1, $stepsize_{normal}$ = 2, $stepsize_{fast}$ = 4. Training in the speed-manipulation experiment was performed with $stepsize_{normal}$. In teleport experiments, the agent was trained on T = 200 trials before experiencing the teleport manipulation. In the speed experiment, the agent was trained on

T = 500 trials before experiencing the speed manipulation. Cached values were clamped during test trials to prevent learning. Agents were simulated with the following parameters: α_{TD} = 0.01, α_{MB} = 0.50, γ = 0.93, k = 0.50. RPEs were extracted on test trials and visualised in fig. 4.

6.7 Mikhael et al. (2022) simulation

Finally, we compared dual-process RPEs with dopamine signals in Mikhael et al. (2022). In the Mikhael et al. (2022) experiment, subjects viewed a VR track akin to Kim et al. (2020) except that the sensory features were progressively darkened on a subset of trials. The experiment further incorporated speed-manipulations.

Environments were formalised with feature-based function approximation. Each state was initially encoded as a one-hot feature vector. To generate state uncertainty, feature vectors were passed through a Gaussian filter parameterised by $\mathcal{N}(s_t, \sigma_t)$. The mean of the Gaussian s_t was always the true-state at time t, while the standard deviation $sigma_t$ was time dependent, and set differently in each experimental condition (see below). Lost probability mass (i.e. mass that was pushed beyond the boundaries of the environment due to Gaussian filtering) was reassigned to the nearest boundary state ensure simplex feature distributions.

In fig. 5C, we simulated a dual-process agent on the VR track experiment in Mikhael et al. (2022). The agent was simulated according to eqs. 5—7, but with value estimates constructed according to eq. (11) and eq. (12) to account for state uncertainty. The agent was tested on a linear track with N = 87 states with a goal state that delivered a reward r = 1.0 at one end of the track. In the bright condition, the standard-deviation in the Gaussian filter was constant at σ_t = 4. In the darkening condition, the standard-deviation was drawn from a rescaled exponential function with the minimum value σ_{min} = 4 and a maximum value σ_{max} = 24, which reproduced the assumptions about state uncertainty during sensory darkening in Mikhael et al. (2022). In the standard-speed condition, the agent moved with the stepsize $stepsize_{std}$ = 1, whereas in the fast-speed condition, it moved with the stepsize $stepsize_{fast}$ = 2. The agent was simulated with the following parameters: α_{TD} = 0.01, α_{MB} = 0.50, γ = 0.93, k = 0.50. It was first trained on the task in the bright, standard-speed condition for T = 150 trials. It then performed one test trial in each combination of brightness (bright-vs-dark) and speed (standard-vs-fast) conditions. The RPE in each state and each test condition was extracted and visualised in fig. 5C.

Data availability

Code to reproduce the results in the manuscript is available at:
https://github.com/lpriestley/dopamine_ramps

5 Acknowledgements

We are grateful to Kris Jensen, Eleanor Spens and Marta Blanco-Pozo for helpful feedback on the manuscript. The work was supported by Wellcome Trust Career Development Award 225926/Z/22/Z. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Additional information

Funding

Funder	Grant reference number	Author
Wellcome Trust (WT)	https://doi.org/10.35802/225926	Thomas Akam

Author ORCID iDs

Luke Priestley: https://orcid.org/0009-0003-9125-7265
Thomas Akam: https://orcid.org/0000-0002-1810-0494

References

Akam T., Walton M. E. (2021) What is dopamine doing in model-based reinforcement learning?. *Current Opinion in Behavioral Sciences* **38**:74-82 https://doi.org/10.1016/j.cobeha.2020.10.010 | PubMed

Akam T., Rodrigues-Vaz I., Marcelo I., Zhang X., Pereira M., Oliveira R. F., Dayan P., Costa R. M. (2021) The Anterior Cingulate Cortex Predicts Future States to Mediate Model-Based Action Selection. *Neuron* **109**:149-163.e7, https://doi.org/10.1016/j.neuron.2020.10.013 | PubMed

Aronov D., Nevers R., Tank D. W. (2017) Mapping of a non-spatial dimension by the hippocam-pal–entorhinal circuit. *Nature* **543**:719-722 https://doi.org/10.1038/nature21692 | PubMed

Babiczky Á., Matyas F. (2022) Molecular characteristics and laminar distribution of prefrontal neurons projecting to the mesolimbic system. *eLife* **11**:e78813 https://doi.org/10.7554/eLife.78813 | PubMed

Balleine B. W., Dickinson A. (1998) Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology* **37**:407-419 https://doi.org/10.1016/S0028-3908(98)00033-1 | PubMed

Beier K. T., Steinberg E. E., DeLoach K. E., Xie S., Miyamichi K., Schwarz L., Gao X. J., Kremer E. J., Malenka R. C., Luo L. (2015) Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic Input-Output Mapping. *Cell* **162**:622-634 https://doi.org/10.1016/j.cell.2015.07.015 | PubMed

Berke J. D. (2018) What does dopamine mean?. *Nature Neuroscience* **21**:787-793 https://doi.org/10.1038/s41593-018-0152-y | PubMed

Blanco-Pozo M., Akam T., Walton M. E. (2024) Dopamine-independent effect of rewards on choices through hidden-state inference. *Nature Neuroscience* **27**:286-297 https://doi.org/10.1038/s41593-023-01542-x | PubMed

Bush D., Barry C., Manson D., Burgess N. (2015) Using Grid Cells for Navigation. *Neuron* **87**:507-520 https://doi.org/10.1016/j.neuron.2015.07.006 | PubMed

Campbell M. G., Ra Y., Chen Z., Xu S., Burrell M., Matias S., Watabe-Uchida M., Uchida N. (2025) A hardwired neural circuit for temporal difference learning. *bioRxiv* 2025.09.18.677203 https://doi.org/10.1101/2025.09.18.677203

Chang L., Tsao D. Y. (2017) The Code for Facial Identity in the Primate Brain. *Cell* **169**:1013-1028.e14, https://doi.org/10.1016/j.cell.2017.05.011 | PubMed

Cohen J. Y., Haesler S., Vong L., Lowell B. B., Uchida N. (2012) Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature* **482**:85-88 https://doi.org/10.1038/nature10754 | PubMed

Constantinescu O., O'Reilly J. X., Behrens T. E. J. (2016) Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**:1464-1468 https://doi.org/10.1126/science.aaf0941 | PubMed

Daw N. D., Niv Y., Dayan P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience* **8**:1704-1711 https://doi.org/10.1038/nn1560 | PubMed

Daw N. D., Gershman S. J., Seymour B., Dayan P., Dolan R. J. (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**:1204-1215 https://doi.org/10.1016/j.neuron.2011.02.027 | PubMed

Dayan P. (1993) Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation* **5**:613-624 https://doi.org/10.1162/neco.1993.5.4.613

Dolan R. J., Dayan P. (2013) Goals and Habits in the Brain. *Neuron* **80**:312-325 https://doi.org/10.1016/j.neuron.2013.09.007 | PubMed

Donnarumma F., Parr T., Friston K., Whittington J., Pezzulo G. (2025) Inferential planning in the frontal cortex. *bioRxiv* 2025.11.26.690672 https://doi.org/10.1101/2025.11.26.690672

Eshel N., Bukwich M., Rao V., Hemmelder V., Tian J., Uchida N. (2015) Arithmetic and local circuitry underlying dopamine prediction errors. *Nature* **525**:243-246 https://doi.org/10.1038/nature14855 | PubMed

Farrell K., Lak A., Saleem A. B. (2022) Midbrain dopamine neurons signal phasic and ramping reward prediction error during goal-directed navigation. *Cell Reports* **41**:111470 https://doi.org/10.1016/j.celrep.2022.111470 | PubMed

Gao L., Liu S., Gou L., Hu Y., Liu Y., Deng L., Ma D., Wang H., Yang Q., Chen Z., *et al.* (2022) Single-neuron projectome of mouse prefrontal cortex. *Nature Neuroscience* **25**:515-529 https://doi.org/10.1038/s41593-022-01041-5 | PubMed

Guru C. Seo, Post R. J., Kullakanda D. S., Schaffer J. A., Warden M. R. (2020) Ramping activity in midbrain dopamine neurons signifies the use of a cognitive map. *bioRxiv* https://doi.org/10.1101/2020.05.21.108886

Hafting T., Fyhn M., Molden S., Moser M.-B., Moser E. I. (2005) Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**:801-806 https://doi.org/10.1038/nature03721 | PubMed

Hamid A., Pettibone J. R., Mabrouk O. S., Hetrick V. L., Schmidt R., Vander Weele C. M., Kennedy R. T., Aragona B. J., Berke J. D. (2016) Mesolimbic dopamine signals the value of work. *Nature Neuroscience* **19**:117-126 https://doi.org/10.1038/nn.4173 | PubMed

Howe M. W., Tierney P. L., Sandberg S. G., Phillips P. E. M., Graybiel A. M. (2013) Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature* **500**:575-579 https://doi.org/10.1038/nature12475 | PubMed

Huang Y., Yaple Z. A., Yu R. (2020) Goal-oriented and habitual decisions: Neural signatures of model-based and model-free learning. *NeuroImage* **215** https://doi.org/10.1016/j.neuroimage.2020.116834 | PubMed

Jensen K. T., Doohan P., Sablé-Meyer M., Reinert S., Baram A., Akam T., Behrens T. E. J. (2026) A mechanistic theory of planning in prefrontal cortex. *eLife* **15**:RP109757 https://doi.org/10.7554/eLife.109757.1

Jones J. L., Esber G. R., McDannald M. A., Gruber A. J., Hernandez A., Mirenzi A., Schoenbaum G. (2012) Orbitofrontal Cortex Supports Behavior and Learning Using Inferred But Not Cached Values. *Science* **338**:953-956 https://doi.org/10.1126/science.1227489 | PubMed

Kato A., Morita K. (2016) Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. *PLOS Computational Biology* **12**:e1005145 https://doi.org/10.1371/journal.pcbi.1005145 | PubMed

Killcross S., Coutureau E. (2003) Coordination of Actions and Habits in the Medial Prefrontal Cortex of Rats. *Cerebral Cortex* **13**:400-408 https://doi.org/10.1093/cercor/13.4.400 | PubMed

Kim H. R., Malik A. N., Mikhael J. G., Bech P., Tsutsui-Kimura I., Sun F., Zhang Y., Li Y., Watabe-Uchida M., Gershman S. J., *et al.* (2020) A Unified Framework for Dopamine Signals across Timescales. *Cell* **183**:1600-1616.e25, https://doi.org/10.1016/j.cell.2020.11.013 | PubMed

Krausz T. A., Comrie A. E., Kahn A. E., Frank L. M., Daw N. D., Berke J. D. (2023) Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron* **111**:3465-3478.e7, https://doi.org/10.1016/j.neuron.2023.07.017 | PubMed

Lee S. W., Shimojo S., O'Doherty J. P. (2014) Neural Computations Underlying Arbitration between Model-Based and Model-free Learning. *Neuron* **81**:687-699 https://doi.org/10.1016/j.neuron.2013.11.028 | PubMed

Liu L., She L., Chen M., Liu T., Lu H. D., Dan Y., Poo M.-m. (2016) Spatial structure of neuronal receptive field in awake monkey secondary visual cortex (V2). *Proceedings of the National Academy of Sciences* **113**:1913-1918 https://doi.org/10.1073/pnas.1525505113 | PubMed

Lloyd K., Dayan P. (2015) Tamping Ramping: Algorithmic, Implementational, and Computational Explanations of Phasic Dopamine Signals in the Accumbens. *PLOS Computational Biology* **11**:e1004622 https://doi.org/10.1371/journal.pcbi.1004622 | PubMed

Mattar M. G., Daw N. D. (2018) Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience* **21**:1609-1617 https://doi.org/10.1038/s41593-018-0232-z | PubMed

Mattar M. G., Lengyel M. (2022) Planning in the brain. *Neuron* **110**:914-934 https://doi.org/10.1016/j.neuron.2021.12.018 | PubMed

Mikhael J. G., Kim H. R., Uchida N., Gershman S. J. (2022) The role of state uncertainty in the dynamics of dopamine. *Current Biology* **32**:1077-1087.e9, https://doi.org/10.1016/j.cub.2022.01.025 | PubMed

Mohebi A., Pettibone J. R., Hamid A. A., Wong J.-M. T., Vinson L. T., Patriarchi T., Tian L., Kennedy R. T., Berke J. D. (2019) Dissociable dopamine dynamics for learning and motivation. *Nature* **570**:65-70 https://doi.org/10.1038/s41586-019-1235-y | PubMed

Montague P., Dayan P., Sejnowski T. (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience* **16**:1936-1947 https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996 | PubMed

Niedringhaus M., West E. A. (2022) Prelimbic cortex neural encoding dynamically tracks expected outcome value. *Physiology & Behavior* **256** https://doi.org/10.1016/j.physbeh.2022.113938 | PubMed

Niv Y. (2013) Dopamine ramps up. *Nature* **500**:533-535 https://doi.org/10.1038/500533a | PubMed

O'Doherty J. P., Dayan P., Friston K., Critchley H., Dolan R. J. (2003) Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron* **38**:329-337 https://doi.org/10.1016/S0896-6273(03)00169-7 | PubMed

Pawlak V., Kerr J. N. D. (2008) Dopamine Receptor Activation Is Required for Corticostriatal Spike-Timing-Dependent Plasticity. *The Journal of Neuroscience* **28**:2435-2446 https://doi.org/10.1523/JNEUROSCI.4402-07.2008 | PubMed

Pessiglione M., Seymour B., Flandin G., Dolan R. J., Frith C. D. (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature* **442**:1042-1045 https://doi.org/10.1038/nature05051 | PubMed

Piray P., Daw N. D. (2021) Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature Communications* **12**:4942 https://doi.org/10.1038/s41467-021-25123-3 | PubMed

Ringstrom T. J., Hasanbeig M., Abate A. (2025) Goal Kernel Planning: Linearly-Solvable Non-Markovian Policies for Logical Tasks with Goal-Conditioned Options. *arXiv* https://doi.org/10.48550/arXiv.2007.02527

Sagiv Y., Akam T., Witten I. B., Daw N. D. (2025) Prioritizing replay when future goals are unknown. *Neuron* **113**:4278-4292 https://doi.org/10.1016/j.neuron.2025.09.021 | PubMed

Samejima K., Ueda Y., Doya K., Kimura M. (2005) Representation of Action-Specific Reward Values in the Striatum. *Science* **310**:1337-1340 https://doi.org/10.1126/science.1115270 | PubMed

Schultz W. (2006) Behavioral Theories and the Neurophysiology of Reward. *Annual Review of Psychology* **57**:87-115 https://doi.org/10.1146/annurev.psych.56.091103.070229 | PubMed

Schultz W., Dayan P., Montague P. R. (1997) A Neural Substrate of Prediction and Reward. *Science* **275**:1593-1599 https://doi.org/10.1126/science.275.5306.1593 | PubMed

Stalnaker T. A., Cooch N. K., McDannald M. A., Liu T.-L., Wied H., Schoenbaum G. (2014) Orbitofrontal neurons infer the value and identity of predicted outcomes. *Nature Communications* **5**:3926 https://doi.org/10.1038/ncomms4926 | PubMed

Steinberg E. E., Keiflin R., Boivin J. R., Witten I. B., Deisseroth K., Janak P. H. (2013) A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience* **16**:966-973 https://doi.org/10.1038/nn.3413 | PubMed

Sutton R. S. (1991) Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin* **2**:160-163 https://doi.org/10.1145/122344.122377

Sutton R. S., Barto A. (2014) *Reinforcement Learning: An Introduction. Adaptive Computation and Machine Learning* (nachdruck edition) Cambridge, Massachusetts: The MIT Press.

Van Der Meer M. A. A. (2009) Covert expectation-of-reward in rat ventral striatum at decision points. *Frontiers in Integrative Neuroscience* **3** https://doi.org/10.3389/neuro.07.001.2009 | PubMed

Wang Q., Ding S.-L., Li Y., Royall J., Feng D., Lesnar P., Graddis N., Naeemi M., Facer B., Ho A., *et al.* (2020) The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. *Cell* **181**:936-953.e20, https://doi.org/10.1016/j.cell.2020.04.007 | PubMed

Whittington J. C., Muller T. H., Mark S., Chen G., Barry C., Burgess N., Behrens T. E. (2020) The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell* **183**:1249-1263.e23, https://doi.org/10.1016/j.cell.2020.10.024 | PubMed

Witten B., Steinberg E. E., Lee S. Y., Davidson T. J., Zalocusky K. A., Brodsky M., Yizhar O., Cho S. L., Gong S., Ramakrishnan C., *et al.* (2011) Recombinase-Driver Rat Lines: Tools, Techniques, and Optogenetic Application to Dopamine-Mediated Reinforcement. *Neuron* **72**:721-733 https://doi.org/10.1016/j.neuron.2011.10.028 | PubMed

Yamins L. K., DiCarlo J. J. (2016) Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience* **19**:356-365 https://doi.org/10.1038/nn.4244 | PubMed

Peer reviews

Reviewer #1 (Public review):

Summary:

This study develops a novel theory to account for various aspects of dopamine signals, particularly dopamine ramps. They propose that dopamine reward prediction error (RPE) signals are generated by a dual-process learning system in which values inferred by a model-based system enter the RPE asymmetrically into the update target but not the prediction (equation 6). The work offers specific, mechanistic explanations of Krausz et al. (2023) and Guru et al. (2020), Kim et al. (2020) by maintaining an RPE interpretation, and presents an alternative to the state-uncertainty account in Mikhael et al. (2022) that doesn't require the asymmetric uncertainty assumption Mikhael needs, using Campbell et al. (2025) in a thoughtful way. The asymmetric-RPE idea is clean and well presented. Overall, this study makes an important contribution to the field.

Strengths:

The theory is relatively simple and intuitive. It addresses a long-standing controversy or mystery in the field of dopamine.

Weaknesses:

(1) The biggest outstanding question is what V_TD does - letting V_MB drive everything would seem to produce much of the same outcomes in the settings discussed here. The discussion suggests that in situations where there is little contribution of the model-based system, the backpropagating bump is a feature (e.g. Amo et al.). It would be interesting to see if this is a true outcome of the model, potentially by varying the arbitration parameter k. This is an interesting alternative account from eligibility trace explanations of the lack of backpropagating bump in some experimental settings.

(2) The model-based accounts are quite simplistic, and this should probably be acknowledged - it does help delineate their contribution, but in the model, only the goal-reward value is updated; everything else is a known computation. Perhaps engage more deeply with Sagiv et al?

(3) The application of Campbell et al. (2025) to push back on Mikhael (lines 253-259) is interesting: if striatum to VTA implements TD via synaptic delays such that V(s_t) is a delayed

copy of V(s_{t+1}), then state uncertainty is necessarily shared between the two terms in the RPE, defeating Mikhael's required asymmetry.

But the same circuit logic creates tension for the dual-process model. It seems they are proposing that the frontal cortex projects V_MB into VTA dopamine neurons (as proposed in 3.1 and the Discussion) and adds to the prediction error derived from the biphasic filtering of value. But the biphasic idea (and data of Campbell et al.) implies that the V(t+1) and -V(t) come from the same source and are proportional. Adding the V_MB term is akin to adding a positive bias, breaking the optimality of the TD error for predicting value and predicting over-learning of cached value. It is worth considering whether V_MB passes through a similar filter - I am not sure if it is fatal if V_MB contributes somewhat to the negative term of the update error.

(4) A few places where the predicate of the conclusion needs more care. The "normative" framing throughout 3.2 and the Discussion is normative conditional on the architecture already including a separate cached system that needs to converge to the true value function and on a system in which the model based is learnt much faster - see comments about learning rate parameter later.

(5) Kim et al. is cited heavily as a data source for Figure 4, but is never engaged with as a theoretical alternative, even though Kim et al. explicitly argued that an appropriate state representation makes standard TD compatible with ramps and the teleport responses. That is, Kim et al. is already a TD account of these phenomena, and doesn't require a second learning system. The introduction and Mikhael discussion treat the field as if the choice were between "dopamine = value" (Hamid, Howe, Mohebi) and dopamine = RPE-with-special-conditions (Mikhael, Kato-Morita), but Kim et al.'s framework is also dopamine = RPE. Two specific places this matters: (i) Figure 4 currently demonstrates that the dual-process model reproduces the Kim teleport results, but Kim et al.'s framework also reproduces them - the figure doesn't distinguish the two, and I am not sure the figure gives this message cleanly. (ii) Kim et al. report that ramps develop with training over days; the manuscript should address whether the dual-process model has an alternative explanation for this, especially given the contrast with the Guru result (ramps diminishing with training over a longer timescale).

(6) The arbitration parameter k is fixed at 0.5 throughout, and the paper acknowledges this is for simplicity, but a supplementary panel sweeping $k \in \{0, 0.2, 0.5, 0.8, 1.0\}$ on the key figures (Figure 1B convergence, Figure 2D ramp dynamics, Figure 3D Krausz updating) would be informative. At k = 0, the model reduces to standard TD; at k = 1, it's effectively V_MB-driven. I think these would be easy to add and help clarify the work this assumption is doing.

(7) Learning-rate asymmetry needs justification. The story relies on $\alpha_{MB} >> \alpha_{TD}$ throughout ($\alpha_{MB} = 0.50$, $\alpha_{TD} = 0.01$ - a 50× ratio). With $\alpha_{MB} = 0.5$, a single rewarded trial moves R[goal] halfway to the new value, which would predict strong dependence of dopamine ramp amplitude on the previous trial's outcome. This is testable in existing data (Krausz et al. should have enough trials to fit the exponential decay constant for trial-history dependence; Guru's swap-session data likewise), and the paper would be strengthened by explicitly deriving and checking that prediction.

(8) α_{MB} is dropped to 0.10 specifically for the Krausz simulation without justification in the text - Why? Either the value should be the same as elsewhere, or the paper should explain why Krausz's task requires slower MB learning. It would be good to check the robustness of the Krausz simulation - the test phase is a single set of three trials (t-2 = omission, t-1 = reward, then t = 50% rewarded) after training on a single set of 500 simulated trials (believe only one random seed is used - given the high alpha, varying this set of simulated trials seems important). Also, do they get the other result in Krausz (t-2 = reward, t-1 = omission, t = 50% rewarded)?

(9) It might be possible to fit the alpha to the Guru and Krausz simulations - this might be informative to show the range over which it varies.

(10) The Kato and Morita account is cited in the introduction but never really discussed again - it would be good to engage with this a bit more in the discussion. The rejection of the value-based accounts seems to rely primarily on Kim et al., where the value and TDRPE accounts differ, but this could be directly acknowledged, rather than absorbing credit for this into their model.

https://doi.org/10.7554/eLife.111458.1.sa2

Reviewer #2 (Public review):

Summary:

This paper offers a novel theoretical account of dopamine ramps. The key idea is that the reward prediction error (putatively signaled by dopamine) uses a partially model-based estimate for future value (the prediction target). Because the model-based value estimate emerges more rapidly than the model-free estimate, it inflates the RPE, and this inflation increases with reward proximity - hence ramps. The authors show that this account can explain many aspects of existing data on dopamine ramps across several different studies.

Strengths:

Overall, I liked this paper. The idea is interesting and plausible. The paper is well-written and clearly argued. The modeling has been done rigorously.

Weaknesses:

My major comments are: (1) it's not always clear which phenomena are uniquely well-explained by this new account vs. earlier accounts; and (2) the limitations of the account are not entirely transparent.

(1) The paper models some of the studies reported by Kim et al (2020). As was already shown in that paper, a standard TD error could explain the results (although a major limitation of that treatment was that it did not model the recursive effect of RPEs on learning, as discussed in the Mikhael paper). It's not clear if there's additional explanatory value provided by this new account, though, of course, it's good to know that those results are captured by the new account. Likewise, Mikhael et al (2022) already offered an account of their data (somewhat more complex than the standard TD model). Again, it's not clear if there's additional explanatory value provided by the new account (and again, it's nice to see that the model can capture these results). Finally, I found myself wondering whether the Guru et al (2020) result couldn't be explained by a more standard TD model (assuming the value function is sufficiently convex). I don't think it's essential that the new account provides additional explanatory value in every case, but I think it's important to convey to readers what's new and what's not, as well as what aspects of the data require particular kinds of mechanisms to explain. It would be really helpful to see the predictions of alternative TD models in order to make this clearer.

(2) The Mikhael model was motivated by the puzzle that ramping is observed in navigation tasks (with sensory cues) but typically not in classical conditioning tasks lacking sensory cues. The correction term, derived from normative considerations, explained this discrepancy. It's not clear to me if/how the new account can explain the discrepancy.

https://doi.org/10.7554/eLife.111458.1.sa1

Reviewer #3 (Public review):

Summary:

This work presents a new hypothesis for why dopamine signals have sometimes been observed to "ramp up" in spatial tasks as rodents approach a location associated with reward. In essence, the hypothesis is that value estimates (i.e., predictions about future rewards) from a model-based system, which may be able to more quickly form such estimates via an inference-like process, can be used to speed up the (relatively slow) learning of such estimates by a model-free system. This is suggested to occur by including the model-based estimate as part of the target towards which model-free estimates are updated in the course of temporal-difference (TD) learning. The early discrepancy between these estimates can be expected to give rise to systematic TD errors - putatively represented in dopaminergic activity - that give rise to dopamine ramps, which are expected to diminish over time as the estimates of both systems converge. The authors show that a model that implements this idea makes predictions about dopamine activity that are a good qualitative match to data from a number of recent experimental studies.

Strengths:

The work suggests a normative account for a phenomenon that has persistently troubled the canonical theory of dopamine function. The account is appealing in its elegance and simplicity, and the authors present compelling evidence that it can capture the empirical observations of key recent papers. Another strength of the account is that it readily suggests avenues for future theory development and experimental test, including what the 'best' target estimate should be at any given time, how rapidly one might expect ramps to develop or diminish, and the neural implementation of the proposed algorithm. This is likely to stimulate further theoretical and experimental work in the field.

Weaknesses:

One aspect of dopamine "ramps" that was troubling from a theoretical standpoint was their apparent persistence over time. Given the authors' prediction that these would disappear over time in a stable environment and the supporting evidence they cite (from Guru et al., 2000), the reader might be left confused about the state of evidence about whether dopamine ramps persist or not. Perhaps relatedly, the issue of how the activity of dopamine cells and dopamine release are related is not discussed, which may be relevant given that early studies (e.g., Howe et al., 2013) used voltammetry to measure extracellular dopamine concentrations.

https://doi.org/10.7554/eLife.111458.1.sa0