Learning with reward prediction errors in a model of the *Drosophila* mushroom body

James E. M. Bennett¹, Andrew Philippides¹, Thomas Nowotny¹

¹Department of Informatics, University of Sussex, UK

Abbreviated title: Reward prediction errors in *Drosophila*

Corresponding author:

James E. M. Bennett

School of Engineering and Informatics

University of Sussex

Chichester 1, Room 002

Falmer

Brighton

BN1 9QJ

Abstract

Effective decision making in a changing environment demands that accurate predictions are learned about decision outcomes. In *Drosophila*, such learning is orchestrated in part by the mushroom body (MB), where dopamine neurons (DANs) signal reinforcing stimuli to modulate plasticity presynaptic to MB output neurons (MBONs). Here, we extend previous MB models, in which DANs signal absolute rewards, proposing instead that DANs signal reward prediction errors (RPEs) by utilising feedback reward predictions from MBONs. We formulate plasticity rules that minimise RPEs, and use simulations to verify that MBONs learn accurate reward predictions. We postulate as yet unobserved connectivity, which not only overcomes limitations in the experimentally constrained model, but also explains additional experimental observations that connect MB physiology to learning. The original, experimentally constrained model and the augmented model capture a broad range of established fly behaviours, and together make five predictions that can be tested using established experimental methods.

Introduction

Expedient decision making benefits from an organism's ability to accurately predict the rewarding and punishing outcomes of each decision, so that it can meaningfully compare the available options and act to bring about the greatest reward. In many scenarios, an organism must learn to associate the valence of each outcome with the sensory cues predicting it. A broadly successful theory of reward learning is the delta rule (1), whereby reward predictions are updated in proportion to reward prediction errors (RPEs): the difference between predicted and received rewards. RPEs are more effective as a learning signal than absolute rewards because RPEs diminish as the prediction becomes more accurate, adding stability to the learning process. In mammals, RPEs are signalled by dopamine neurons (DANs) in the ventral tegmental area and substantia nigra, enabling the brain to implement approximations to the delta rule (2; 3). In *Drosophila melanogaster*, DANs that project to the mushroom body (MB) (Fig. 1a) also provide reward modulated signals that are required for associative learning (4). However, to date, MB DAN activity is typically interpreted as signalling absolute reinforcement, and direct evidence for RPE signals is lacking. Here, we build a novel computational model of the MB, taking advantage of precise anatomical and functional data from recent experiments, and provide a circuit-level description for delta rule learning in the MB, whereby MB DANs compute RPEs.

Consistent with its role in associative learning, the MB is organised into lateral and medial lobes of neuropil that, in addition to DANs, are innervated by the axons of sensory encoding Kenyon cells (KCs), and the dendrites of MB output neurons (MBONs), which modulate behaviour (Fig. 1b). Current models of MB function posit that the MB lobes encode different valences of reward information and actions (5–11). Most DANs in the PAM (protocerebral anterior medial) cluster (called D_+ in the model presented here) are activated by rewards, or positive reinforcement (R_+), and their activation results in depression at synapses between coactive KCs (K) and MBONs that are thought to induce avoidance behaviours (M_-). DANs in the PPL1 (protocerebral posterior lateral 1) cluster (D_-) are activated by punishments, i.e. negative reinforcement (R_-), and their activation results in depression at synapses between coactive KCs and MBONs that induce approach behaviours (M_+). A fly can therefore learn to approach rewarding cues or avoid punishing cues as a result of synaptic depression at KC inputs to avoidance or approach MBONs, respectively.

To date, there are only indirect lines of evidence for RPE signals in MB DANs. DAN activity is modulated by feedforward reward signals, but some DANs also receive excitatory feedback from MBONs (12–15), and it is likely this extends to all MBONs whose axons are proximal to DAN dendrites (16). MBON activity is, in turn, modulated by DANs via plasticity at KC→MBON synapses (10; 17; 18). We interpret the difference between approach and avoidance MBON firing rates

as a reward prediction that motivates behaviour, consistent with the observation that behavioural valence scales with the difference between approach and avoidance MBON firing rates (9). As such, DANs that integrate feedforward reward signals and feedback reward predictions from MBONs are primed to signal RPEs for learning. To the best of our knowledge, these latter two features have yet to be incorporated in computational models of the MB (19–21).

Here, we formulate a reduced computational model of the MB circuitry described above to demonstrate how DANs may compute RPEs, then, hypothesising that learning minimises RPEs, derive a plasticity rule for KC→MBON synapses, and verify in simulations that it yields accurate reward predictions. We identify limitations to the model that impose an upper bound to reward prediction magnitudes, and demonstrate how as yet unobserved connections between DANs, KCs and MBONs could circumvent this limitation. Introducing these additional connections yields testable predictions for future experiments as well as explaining a broader range of existing experimental observations that connect DAN and MBON stimulus responses to learning. Lastly, we show that both incarnations of the model – with and without additional connections – capture a wide range of experimental observations in *Drosophila*, whereby conditioned approach and avoidance behaviours are compromised by genetically targeted manipulation of MBON and DAN activity. Different behavioural outcomes in the two models for particular genetic manipulations provide further strong experimental predictions.

Results

A computational model of the mushroom body

The MB lobes comprise multiple compartments, each innervated by a different set of MBONs and DANs (Fig. 1b), and each encoding memories for different forms of reward (22), with different longevities (23), and for different stages of memory formation (24). Nevertheless, compartments appear to contribute to learning by similar mechanisms (10; 18; 25), and it is reasonable to assume that the process of learning reward predictions is similar for different forms of reward. We therefore reduce the multicompartmental MB into two compartments, and assign a single, rate-based unit to each class of MBON and DAN (colour-coded in Fig. 1B-C). KCs, however, are modelled as a population, in which each sensory cue selectively activates a unique subset of 10 units. Given that activity in approach and avoidance MBONs – denoted M_+ and M_- in our model – respectively bias flies to approach or avoid a cue, i , we interpret the difference in their firing rates, $\hat{m}^i = m_+^i - m_-^i$, as the fly's reward prediction for that cue.

For the purpose of this work, we assume that the MB has only a single objective: to form reward predictions that are as accurate as possible, i.e. that minimise the RPE. We do this within a multiple-alternative forced choice (MAFC) paradigm

(Fig. 1D; also known as a multi-armed bandit) in which a fly is exposed to one or more sensory cues in a given trial, and is forced to choose one. The fly then receives a reinforcement signal, $\hat{r}^i = r_+^i - r_-^i$, which has both positive and negative components (coming from sources R_+ and R_-), and which is specific to the chosen cue. Over several trials, the fly must learn to predict the rewards for each cue, and use these predictions to reliably choose the most rewarding cue. We formalise this objective with a cost function

$$C^{\text{RPE}} = \frac{1}{2} \sum_i (\hat{r}^i - \hat{m}^i)^2, \quad (1)$$

where the sum is over all cues, i . By performing gradient descent on C^{RPE} , we derived a learning rule, $\mathcal{P}^{\text{RPE}}$ (full derivation in Methods: Synaptic plasticity):

$$\mathcal{P}_{\pm}^{\text{RPE}} = \frac{\eta}{2} \mathbf{k} (d_{\pm} - d_{\mp}). \quad (2)$$

which minimises the cost when used to update synaptic weights according to $\mathbf{w}_{\pm}(t+1) = \mathbf{w}_{\pm}(t) + \mathcal{P}_{\pm}^{\text{RPE}}$. We use subscripts ‘ $\pm$ ’ to denote the valence of the neuron: if $+$ ($-$) is considered in $\pm$, then $\mp$ refers to $-$ ($+$), and *vice versa*. As such, $d_{\pm}$ refers to the firing rate of either D_+ or D_- . The learning rate, η , must be small (see Methods: Synaptic plasticity) to allow the plasticity rule to average over multiple stimuli as well as stochasticity in the reward schedule (see Methods: Reward schedule). Note that a single DAN, $D_{\pm}$, only has access to half of the reward and reward prediction information, and by itself does not compute the full RPE. However, the difference between D_+ and D_- firing rates does yield the full RPE (see Methods: DAN firing rates):

$$d_+^i - d_-^i = \hat{r}^i - \hat{m}^i. \quad (3)$$

Three features of Eq. 2 are worth highlighting here. First, elevations in $d_{\pm}$ increase the net amount of synaptic depression at active synapses that impinge on $M_{\mp}$, which encodes the opposite valence to $D_{\pm}$, in agreement with experimental data (10; 18; 25). Second, the postsynaptic MBON firing rate is not a factor in the plasticity rule, unlike in reward-modulated Hebbian rules (26), yet nevertheless in accordance with experiments (18). Third, Eq. 2 requires that synapses receive dopamine signals from both D_+ and D_- , conflicting with current experimental findings in which appetitive DANs only modulate plasticity at avoidance MBONS, and similarly for aversive DANs and approach MBONS (10; 17; 18; 22; 27; 28). To accommodate the experimental data, we consider an alternative cost function, $C_{\pm}^{\text{VS}}$, that satisfies the valence specificity of DANs:

$$C_{\pm}^{\text{VS}} = \frac{1}{2} \sum_i (r_{\pm}^i + m_{\mp}^i)^2. \quad (4)$$

Gradient descent on $C_{\pm}^{\text{VS}}$ yields the corresponding valence-specific plasticity rule:

$$\mathcal{P}_{\pm}^{\text{VS}} = \eta \mathbf{k} (\mathbf{w}_{\mathbf{k}}^T \mathbf{k} - d_{\mp}), \quad (5)$$

where $w_K^T k$ models the direct excitatory current from KCs to DANs (Methods, Eq. 11). As required, Eq. 5 maintains the relationship between increased DAN activity and enhanced synaptic depression. We refer to models that utilise this plasticity rule as valence-specific (VS) models. In our simulations, we update only those synaptic weights that correspond to the chosen cue, as only the reward for this cue would be experienced by a fly (see Methods, Eq. 21).

Restricted learning in the valence-specific model

Eq. 5 exposes a problem for learning according to our assumed objective in the VS model. The problem arises because $D_{\pm}$ receives only excitatory inputs. Thus, whenever a cue is present, KC inputs (29) prescribe $D_{\pm}$ with a minimum, cue-specific firing rate, $d_{\pm}^i \geq w_K^T k^i$. As such, synapses will be depressed ($\mathcal{P}_{\mp}^{VS} < 0$) whenever $r_{\pm} > 0$ or $m_{\mp} > 0$. Once $w_K^T k^i = 0$, the VS model can no longer learn the valence of cue i as synaptic weights cannot become negative. Eventually, reward predictions for all cues become equal with $\hat{m}^i = 0$, such that choices become random (Supplementary Fig. 1a-b). In this case, D_+ and D_- firing rates become equal to the positive and negative rewards, respectively, such that RPEs equal the net rewards (Supplementary Fig. 1c-d).

A heuristic solution is to add a constant source of potentiation, which acts to restore synaptic weights to a constant, non-zero value. We therefore replace $w_K^T k$ in Eq. 5 with a constant, free parameter, λ :

$$\mathcal{P}_{\pm}^{VS\lambda} = \eta k (\lambda - d_{\mp}). \quad (6)$$

By setting $\lambda > |r_+ - r_-| + w_K^T k$, $\mathcal{P}_{\pm}^{VS\lambda}$ can take both positive and negative values, preventing synaptic weights from being held at zero. This defines a new baseline firing rate for $D_{\pm}$ that is greater than $w_K^T k$. Hereafter, we refer to the VS model with plasticity governed by $\mathcal{P}_{\pm}^{VS\lambda}$ as the VS λ model.

The VS λ model provides only a partial solution, as it is restricted by an upper bound to the magnitude of reward predictions that can be learned: $|\hat{m}|_{\max} = \max(0, \lambda - w_K^T k)$. In addition to increasing λ , $|\hat{m}|_{\max}$ may be increased by reducing KC→DAN synaptic transmission. In Fig. 2a, we set $w_K = \gamma \mathbf{1}$, with $\mathbf{1}$ a vector of ones, and show reward predictions for several values of γ , with $\lambda = 11.5$ (corresponding DAN and MBON firing rates are in Supplementary Fig. 2). The upper bound is reached when w_+ or w_- , and thus the corresponding MBON firing rates, go to zero (an example when $\gamma = 1$ is shown in Fig. 2b). These results appear to contradict recent experimental work in which learning was impaired, rather than enhanced, by blocking KC→DAN synaptic transmission (29).

In the VS λ model, DAN firing rates begin to exhibit RPE signals. A sudden increase in positive reinforcements, for example at trial 20 in Fig. 2c, results in a sudden increase in d_+ , which then decays as the excitatory feedback from M_- diminishes as a result of synaptic depression in w_- (Fig. 2c-d). Similarly, sudden

decrements in positive reinforcements, for example at trial 80, are signalled by reductions in d_+ . However, when the reward magnitude exceeds the upper bound, as in trials 40-60 and 120-140 in Fig. 2a-d, $D_{\pm}$ exhibits sustained elevations in firing rate from baseline by an amount $\max(0, r_{\pm} - |\hat{m}|_{\max})$ (Fig. 2c, Supplementary Fig. 2). This constitutes a major prediction from our model.

A mushroom body circuit with unbounded learning

In the VS λ model, excitatory reward signals can only be partially offset by decrements to w_+ and w_- . Thus, to overcome the upper bound to reward predictions, a source of inhibition is required. A candidate solution is a circuit in which positive reinforcements, R_+ , inhibit D_- , and similarly, R_- inhibits D_+ (illustrated in Fig. 3a). Such inhibitory reinforcement signals have been observed in the $\gamma 2$, $\gamma 3$, $\gamma 4$ and $\gamma 5$ compartments of the MB (17; 30). Using the derived plasticity rule, $\mathcal{P}_{\pm}^{\text{VS}}$ in Eq. 5, this circuit learns accurate reward predictions with no upper bound to the reward prediction magnitude (Supplementary Fig. 3a). Hereafter, we refer to the VS model with unbounded learning as the VSu model. Learning is now possible because, when the synaptic weights $w_{\pm}$ are weak, or when $D_{\mp}$ is inhibited, Eq. 5 specifies that $w_k^T \mathbf{k} - d_{\mp} > 0$, i.e. synaptic weights will potentiate until the feedback excitation equals the reinforcement-induced feedforward inhibition. Similarly, synapses are depressed in the absence of reinforcement because the excitatory feedback from $M_{\pm}$ to $D_{\mp}$ ensures that $w_k^T \mathbf{k} - d_{\mp} < 0$ (Supplementary Fig. 3b). Consequently, step changes in reinforcement yield RPE signals in $D_{\mp}$ that always decay to a baseline set by $w_k^T \mathbf{k}$ (Supplementary Fig. 3c-d). Despite the prevalence in reports of long term synaptic depression in the MB, there exist several lines of evidence for potentiation as well (10; 11; 14; 31). However, when reinforcement signals are inhibitory, D_+ , for example, is activated only after decrements to r_- , but not following increments to r_+ (similarly for D_-), counter to the experimental classification of DANs as appetitive (or aversive) (6–8; 32).

We therefore asked whether other variations of the VSu model could learn without an upper bound, and identified three criteria (tabulated in Supplementary Table 1) that must be satisfied to achieve this: *i*) learning must be effective, such that positive reinforcement either potentiates excitation of approach behaviours (inhibition of avoidance), or depresses inhibition of approach behaviours (excitation of avoidance), and similarly for negative reinforcement; *ii*) learning must be stable, such that excitatory reinforcement signals are offset via learning, either by synaptic depression of feedback excitation, or by potentiation of feedback inhibition, and similarly for inhibitory reinforcement signals; *iii*) to be unbounded, learning must involve synaptic potentiation, whether reinforcement signals excite DANs that induce potentiation, or inhibit DANs that induce depression. By following these criteria, we identified a dual version of the VSu circuit in Fig. 3a, which is illustrated in Fig. 3b. In this circuit, R_+ excites D_+ and R_- excites D_- . However, DANs induce synaptic potentiation when activated above baseline, while M_+ and M_- are inhibitory, so are interpreted as inducing avoidance and approach behaviours, re-

spectively. Despite their different configurations, reward predictions are identical in each of the dual MB circuits (data not shown).

Neither dual model, by itself, captures all of the experimentally established anatomical and physiological properties of the MB. However, by combining them into one (Fig. 3c), the new model becomes consistent with the circuit properties observed in experiments, but necessitates additional features, which constitute major predictions. First, synapses impinging on either approach or avoidance MBONs must undergo plasticity that is modulated by DANs of both valences. Second, DANs receive both positive and negative reward signals, which are either excitatory or inhibitory, depending on the valences of the reinforcement and the DAN. Third, MBONs provide feedback to DANs of the same valence via inhibitory interneurons, which we propose innervate areas targeted by MBON axons and DAN dendrites (16). Each DAN in this hybrid model now has access to the full reward, $\hat{r}$, and the full reward prediction, $\hat{m}$, or $-\hat{r}$ and $-\hat{m}$, depending on the valence of the DAN. Thus, it is now reasonable to utilise the original plasticity rule, $\mathcal{P}_{\pm}^{\text{RPE}}$ in Eq. 2, which was derived with the objective to minimise the full RPE. As such, DANs facilitate synaptic depression if the DAN and MBON have opposite valences, or facilitate potentiation if they have the same valence. We refer to this new model, under the control of $\mathcal{P}_{\pm}^{\text{RPE}}$, as the mixed valence (MV) model.

Fig. 3d demonstrates that the MV model accurately tracks changing rewards, just as with the dual versions of the VSu model. However, a number of differences from the VSu models can also be seen. First, changing reward predictions are manifest by changes in the firing rates of both M_+ and M_- (Fig. 3e). Moreover, when $w_{\pm}$ reach zero, the changes in $w_{\mp}$ compensate, resulting in larger changes in the firing rate of $M_{\mp}$, as seen between trials 40-60 in Fig. 3e. Second, DANs no longer encode prediction errors exclusively in either positive or negative reinforcers, but their activity is instead correlated with the sign of the full RPE: d_+ and d_- increase with positive and negative RPEs, respectively, and decrease with negative and positive RPEs (Fig. 3f-g). Third, blocking KC→DAN synaptic transmission slows down learning, but does not abolish it entirely (Fig. 3d). With input from KCs blocked, the baseline firing rate of $D_{\pm}$ is zero, and because any given RPE excites one DAN type and inhibits the other, only one of either D_+ or D_- can signal the RPE, reducing the magnitude of $d_{\pm} - d_{\mp}$ in Eq. 2, and therefore the speed of learning (Supplementary Fig. 4). To avoid any slowing down to learning, $w_k^T k$ must be greater than or equal to the RPE. This may explain the 25% reduction in learning performance seen by Cervantes-Sandoval et al. (29).

Decision making in a multiple-alternative forced choice task

We next tested the VS λ and MV models on a task with multiple cues from which to choose. Choices are made using the softmax function (Eq. 9), such that the model more reliably chooses one cue over another when cue-specific reward predictions are more dissimilar. Throughout the task, the cue-specific reinforcements

slowly change (see example reward schedules in Fig. 4a-b), and the model must continually update reward predictions, according to its plasticity rule, in order to choose the most positively reinforcing cues as possible.

In a task with two alternatives, switches in cue choice almost always occur after the actual switch in the reinforcement schedule because of the slow learning rate and the probabilistic nature of decision making (Fig. 4a). The model continues to choose the more rewarding cues when there are as many as 200 (Supplementary Fig. 5a; Fig. 4b shows an example simulation with 5 cues). When there are few cues, the mean reward per trial (RPT) obtained increases with the number of cues, as the number of cues providing a large reward increases. The RPT then decreases monotonically with the number of cues as the model is less able to maintain accurate predictions for all cues. Despite this latter degradation in performance, the VS_{λ} and MV models are only marginally outperformed by a model with perfect plasticity, whereby reward predictions for the chosen cue are set to equal the last obtained reward. Furthermore, when noise is added to the reinforcement schedule, the performance of the perfect plasticity model drops below that of the other models, for which slow learning helps to average over noisy observations (Supplementary Fig. 5b). The model suffers no noticeable decrement in performance when KC responses to different cues overlap, e.g. when a random 5% of 2000 KCs are assigned to each cue (Supplementary Fig. 5a,c-e).

Both models capture learned fly behaviours in a variety of conditioning experiments

To determine how well the VS_{λ} and the MV models capture decision making in flies, we applied them to an experimental paradigm (illustrated in Fig. 5a) in which flies are conditioned to approach or avoid one of two odours. In each experiment, flies undergo a training stage, during which they are exposed to a conditioned stimulus (CS+) concomitantly with an unconditioned stimulus (US), for example sugar (appetitive training) or electric shock (aversive training). Flies are next exposed to a different stimulus (CS-) without any US. Following training, flies are tested for their behavioural valence with respect to the two odours. The CS+ and CS- are released at opposite ends of a tube. Flies are free to approach or avoid the stimuli by walking towards one end of the tube or the other. In our model, we do not simulate the spatial extent of the tube, nor specific fly actions, but model choice behaviour in a simple manner by applying the softmax function to the current reward predictions.

In addition to these control experiments, we simulated a variety of interventions frequently used in experiments (Fig. 5a-c). These experiments are determined by four features: 1) US valence (Fig. 5a): appetitive, aversive, or neutral; 2) intervention type (Fig. 5c): inhibition of neuronal output, e.g. by expression of *shibire*, or activation, e.g. by expression of *dTrpA1*, both of which are controlled by temperature; 3) the intervention schedule (Fig. 5b): during the CS+ only, throughout

CS+ and CS−, during test only, or throughout all stages; 4) the target neuron (Fig. 5c): either M₊, M_−, D₊, or D_−. Further details of these simulations are provided in Methods: Experimental data and model comparisons.

We compared the models to behavioural results from 439 experiments (including 235 controls), which tested 27 unique combinations of the above four parameters in 14 previous studies (7–13; 22; 23; 27; 30; 31; 33; 34) (data provided in Supplementary Table 2). The source and experimental details of each experimental intervention used here is listed in Supplementary Table 2. In Fig. 5d-e, we plot a test statistic, Δ_f , that compares behavioural performance indices (PIs) between a specific intervention experiment and its corresponding control, where the PI is +1 if all flies approached the CS₊, and −1 if all flies approached the CS_−. For example, a positive Δ_f indicates that more flies approached the CS₊ in the intervention than in the control experiment. Interventions in both models correspond well with those in the experiments: Δ_f from the VS λ model and experiments are correlated with $R = 0.65$, and Δ_f from the MV model and experiments are correlated with $R = 0.63$ ($p < 0.0001$ for both models).

Four cases of inhibitory interventions exemplify the correspondence of both the VS λ and MV model with experiments, and are highlighted in Fig. 5d-e (light green, purple, blue and orange rings). Also highlighted are two examples of excitatory interventions, in which artificial stimulation of either D₊ or M_− during CS₊ exposure, without any US, was used to induce an appetitive memory and approach behaviour. The two models do not always yield the same behavioural result. The example highlighted in dark blue, in which M₊ was inhibited throughout appetitive training but not during the test, shows that this intervention had little effect in the MV model, in agreement with experiments (31), but resulted in a strong reduction in the appetitiveness of the CS₊ in the VS λ model ($\Delta_f \approx -4.5$). In the Supplementary Text, we analyse the underlying synaptic weight dynamics that lead to this difference in model behaviours. The analyses show that not only does this intervention amplify the difference between CS₊ and CS_− reward predictions in the MV model, it also results in faster memory decay in the VS λ model. Hence, the preference for the CS₊ is maintained in the MV model, but is diminished in the VS λ model.

Discussion

Overview

Successful decision making relies on the ability to accurately predict, and thus reliably compare, the outcomes of choices that are available to an agent. The delta rule, developed by Rescorla & Wagner (1), which updates beliefs in proportion to a prediction error, provides a method to learn accurate predictions. In this work, we have investigated the hypothesis that, in *Drosophila melanogaster*, the

MB implements the delta rule. We posit that approach and avoidance MBONs together encode reward predictions, and that feedback from MBONs to DANs, if subtracted from feedforward reinforcement signals, endows DANs with the ability to compute RPEs, which are used to modulate synaptic plasticity. We formulated a plasticity rule that minimises RPEs, and verified the effectiveness of the rule in simulations of multiple-alternative forced choice tasks. We demonstrated how the established valence-specific circuitry of the MB restricted the learned reward predictions to within a given range, and postulated additional anatomical features that could overcome this restriction. We have thus presented two MB models that yield RPEs in DAN activity and that learn accurate reward predictions: *i*) the VS_{λ} model, in which plasticity incorporates a constant source of synaptic potentiation; *ii*) the MV model, in which we propose mixed-valence connectivity between DANs, MBONs and $KC \rightarrow MBON$ synapses. Both the VS_{λ} and the MV models receive equally good support from behavioural experiments in which different genetic interventions impaired learning, while the MV model provides a mechanistic account for a greater variety of physiological changes that occur in individual neurons after learning. It is plausible, and can be beneficial, for both the VS_{λ} and MV models operate in parallel in the MB, as separately learning positive and negative aspects of decision outcomes, if they arise from independent sources, is important for context-dependent modulation of behaviour. Such learning has been proposed for the mammalian basal ganglia (35).

Predictions

The models yield predictions that can be tested using established experimental protocols. Below, we specify which model supports each prediction.

Prediction 1 – both models: DAN responses to the unconditioned stimulus (US) should decay towards a baseline over successive CS^{+} -US pairings, as a result of the learned changes in MBON firing rates. To the best of our knowledge, only one study has measured DAN responses throughout learning in *Drosophila* (36). Consistent with DAN responses in our model, Dylla et al. (36) reported such decaying responses in DANs in the γ - and β' -lobes during paired CS^{+} and US stimulation. However, they reported similar decaying responses when the CS^{+} and US were unpaired (separated by 90 seconds) that were not significantly different from the paired condition. The authors concluded that DANs do not exhibit RPEs, and that the decaying DAN responses were a result of non-associative plasticity. An alternative interpretation is that a 90 second gap between CS^{+} and US does not induce DAN responses that are significantly different from the paired condition, and that additional processes prevent the behavioural expression of learning. Ultimately, the evidence for either effect is insufficient. Moreover, Dylla et al. (36) observed increases in DAN responses to the CS^{+} during training. This is consistent with the temporal difference learning algorithm (37; 38), which is an extension of the Rescorla-Wagner rule.

Prediction 2 – VS λ model: a sufficiently large reward will yield a sustained DAN response. This feature may be tested by conditioning *Drosophila* with a stepped electric shock schedule. DAN responses may initially exhibit RPEs, which should transition to sustained responses as the voltage increases. The absence of sustained responses could imply that λ adapts to the reward history.

Prediction 3 – MV model: the valence of a DAN is defined by its response to RPEs, rather than to rewards *per se*. Thus, DANs previously thought to be excited by positive (negative) reinforcement are in fact excited by positive (negative) RPEs. For example, a reduction in electric shock magnitude, after an initial period of training, would elicit an excitatory (inhibitory) response in appetitive (aversive) DANs. Felsenberg et al. (13; 14) provide indirect evidence for this. The authors trained flies on a CS+, then re-exposed the fly to the CS+ without the US. For an appetitive (aversive) US, CS+ re-exposure would have yielded a negative (positive) RPE. By blocking synaptic transmission from aversive (appetitive) DANs during CS+ re-exposure, the authors prevented the extinction of learned approach (avoidance).

Prediction 4 – MV model: learning is mediated by simultaneous plasticity at both approach and avoidance MBON inputs. Appetitive conditioning does indeed potentiate responses in MB-V3/ α 3 and MVP2/ γ 1-pedc approach MBONs (11; 31), and depress responses in M4/ β '/ β '2mp and M6/ γ 5/ β '2a avoidance MBONs (10). Similarly, removal of an expected aversive stimulus, which constitutes a positive RPE, depresses M6/ γ 5/ β '2a avoidance MBONs (14). In addition, aversive conditioning depresses responses in MPV2/ γ 1-pedc MB-V2/ α 2/ α '2 approach MBONs (18; 25), and potentiates responses in M4/ β '/ β '2mp and M6/ γ 5/ β '2a avoidance MBONs (10; 24). However, the potentiation of M4/ β ' and M6 MBONs is at least partially a result of depressed feedforward inhibition from the MVP2 MBON (11; 14). To the best of our knowledge, simultaneous changes in approach and avoidance MBON activity has not yet been observed. A consequence of this coordinated plasticity is that, if plasticity onto one MBON type is blocked, plasticity at the other MBON type should compensate.

Prediction 5 – MV model: DANs of both valence modulate plasticity at MBONs of a single valence. In contrast, anatomical and functional experimental data suggest that, in each MB compartment, the DANs and MBONs have opposite valences (16; 39). However, the GAL4 lines used to label DANs in the PAM cluster often include as many as 20-30 cells each, and it has not yet been determined whether all labelled DANs exhibit the same valence preference. Similarly, the valence encoded by MBONs is not always obvious. In (9), for example, it is not clear whether optogenetically activated MBONs biased flies to approach the light stimulus, or to exhibit no-go behaviour that kept them within the light. In the absence of anatomical evidence for the MV model, it may be elaborated so that either *i)* each compartment contains DANs of both valences, or *ii)* each compartment contains MBONs of both valences (Supplementary Fig. 6). In larval *Drosophila*, there are several examples of cross-compartmental DANs and MBONs (40), but a

full account of the valence encoded by these neurons is yet to be provided.

Previous mushroom body models

We are not the first to present a MB model that learns different rewards for multiple cues for decision making (19–21). However, these models utilise bounded synapses to prevent reward predictions from growing indefinitely with continued reward experience. Thus, given enough training, these models would not differentiate between two cues that were associated with rewards of the same sign, but different magnitudes. Our model builds upon these studies by incorporating feedback from MBONs to DANs, which allows KC→MBON synapses to accurately encode the reward magnitude and sign with stable fixed points that are reached when the RPE signalled by DANs decays to zero.

Model limitations

Central to this work is the assumption that the MB has only a single objective: to minimise the RPE. In reality, an organism must satisfy multiple objectives that may be mutually opposed. In *Drosophila*, anatomically segregated DANs in the γ -lobe encode water rewards, sugar rewards, and motor activity (7; 8; 17; 22), suggesting that *Drosophilae* do indeed learn to satisfy multiple objectives. Multi-objective optimisation is a challenging problem, and goes beyond the scope of this work. Nevertheless, for many objectives, the principle that accurate predictions aid decision making, which forms the basis of this work, still applies.

For simplicity, our simulations compress all events within a trial to a single point in time, and are therefore unable to address some time-dependent features of learning. For example, activating DANs either before or after cue exposure can induce memories with opposite valences (23; 41). Nor have we addressed the credit assignment problem: how to associate a cue with reinforcement when they do not occur simultaneously. A candidate solution is temporal difference (TD) learning (37; 38), whereby reinforcement information is back-propagated in time to all cues that predict it. While DAN responses in the MB hint at TD learning(36), it is not yet clear how the MB circuitry could implement it. An alternative solution is an eligibility trace (38; 42), which enables synaptic weights to be updated upon reinforcement even after presynaptic activity has ceased.

Lastly, our work here addresses memory acquisition, but not memory consolidation, which is supported by distinct circuits within the MB (43). Incorporating memory stabilising mechanisms may help to better align our simulations of genetic interventions with fly behaviour in conditioning experiments.

Summary

We have developed a model of the MB that goes beyond previous models by incorporating feedback from MBONs to DANs, and shown how such a MB circuit can learn accurate reward predictions through DAN mediated RPE signals. The model provides a basis for understanding a broad range of behavioural experiments, and reveals limitations to learning given the anatomical data currently available from the MB. Those limitations may be overcome with additional connectivity between DANs, MBONs and KCs, which provide five strong predictions from our work.

Methods

Experimental paradigm

In all but the last results section, we apply our model to a multi-armed bandit paradigm (38; 44) comprising a sequence of trials, in which the model is forced to choose between a number of cues, each cue being associated with its own reinforcement schedule. In each trial, the reinforcement signal may have either positive valence (reward) or negative valence (punishment, or negative reward), which changes over trials. Initially, the fly is naive to the cue-specific rewards. Thus, in order to reliably choose the most rewarding cue, it must learn, over successive trials, to accurately predict the rewards for each cue. Individual trials comprise three stages in the following order (illustrated in Fig. 1e): *i*) the model is exposed to and computes reward predictions for all cues; *ii*) a choice probability is assigned to each cue using a softmax function (described below), with the largest probability assigned to the cue that predicts the largest reward; *iii*) a single cue is chosen probabilistically, according to the choice probabilities, and the model receives reinforcement with magnitude r_+ (positive reward) or r_- (negative reward). The fly uses this reinforcement signal to update its prediction of the cue-specific reward.

Simulations

Connectivity and synaptic weights

KC→*MBON*

KCs (K in Fig. 1b) constitute the sensory inputs (described below) in our models. Sensory information is transmitted from the KCs, of which there are N_K , to two MBONs, M_+ and M_- , through excitatory, feedforward synapses. For simplicity, we use a subscript x to label the valence of a given sign (e.g. $+$, for approach MBONs), and use y to label the opposite valence (e.g. in this case, $-$, for avoidance MBONs). K_i synapses onto $M_{\pm}$ with a synaptic weight $w_{\pm i}$, which

is initialised with $w_{\pm i} = 0.1\xi_{\pm i}$ for each run of the model, where $\xi_{\pm i}$ is a uniform random variable in the range 0-1.

KC→*DAN*

KCs drive excitatory responses in DANs from the PPL1 cluster (29). In our model, we assume that KCs also provide input to appetitive DANs in the PAM cluster. Thus, K_i drives $D_{\pm}$ through unmodifiable, excitatory synapses with weights, $w_K = \gamma \mathbf{1}$, where $\mathbf{1} = [1, 1, \dots, 1]^T$ is a vector of ones of length N_K .

MBON→*DAN*

MBONs provide excitatory feedback to their respective DANs (12–14). In both the valence-specific (VS) and mixed-valence (MV) models, $M_{\pm}$ synapses onto $D_{\pm}$ with unit synaptic weight. In the mixed-valence (MV) model, $M_{\pm}$ also provides inhibitory feedback to $D_{\pm}$ via an inhibitory interneuron, but we do not model the interneuron explicitly. Thus, we describe the feedback weight simply as $w_M = 1$, and specify whether the input is excitatory or inhibitory in the firing rate equation for $D_{\pm}$ (Eqs. 11 and 12).

Inputs and KC sensory representation

Projection neurons from the antennal lobe and optic lobes provide a substantial majority of inputs to Kenyon cells (KCs) in the MB. These inputs carry olfactory and visual information and, together with recurrent inhibition from the anterior paired lateral neuron, drive a sparse representation of sensory information in approximately 5-10% of the KCs (45–47). For simplicity, we bypass the computations performed in nuclei upstream of the KCs, and assign a unique population of 10 KCs to each cue. Thus, for N_c cues, we simulate $N_K = 10N_c$ KCs. Each KC is always activated by its assigned cue, and each active KC, j , is given the same firing rate, $k_j = 1$ Hz. In a subset of simulations used for Fig. 5d, we simulate 2000 KCs, where each KC is assigned to a cue with probability $p = 0.05$, so that 5% of KCs, on average, are active for a given cue. In these simulations, we normalised the total KC firing rates for each cue, i , such that $\sum_j k_j^i = 10$ Hz. This ensured that the multiplicative effect of KC firing rates on the speed of learning (Eqs. 2 and 5) does not confound the interpretation of our results.

MBON firing rates and reward predictions

Neurons are modelled as linear-nonlinear (LN) units that output a firing rate, y , equal to the rectified linear sum of their inputs, x :

$$y = f\left(\sum_j w_j x_j\right) \quad (7)$$

where $f(z) = \max(0, z)$ is the rectifying nonlinearity. Eq. 7 can be written more concisely in vector notation: $y = f(\mathbf{w}^T \mathbf{x})$, where $\mathbf{w}^T = [w_1, \dots, w_N]$ for N presynaptic neurons, and superscript T denotes the transpose. Throughout this text,

bold fonts denote vectors.

At the beginning of each trial, MBON firing rates, and thus reward predictions, are computed for each cue. The firing rate, $m_{\pm}^i$, of MBON $M_{\pm}$, signals the amount of positive (or negative) reward associated with a given cue, labeled i , according to

$$m_{\pm}^i = f(\mathbf{w}_{\pm}^T \mathbf{k}^i), \quad (8)$$

where $\mathbf{k}^i$ is the vector of KC responses to stimulus i , and $\mathbf{w}_{\pm}$ are plastic, excitatory synaptic weights. The net reward predicted by sensory cue i is then determined by $\hat{m}^i = m_{+}^i - m_{-}^i$.

Decision making

In each trial, reward predictions for all cues are compared, and the model is forced to decide which cue should be chosen. Decisions are made probabilistically using a softmax function, $p(i)$, which specifies the probability of choosing cue i as a function of the differences between its reward prediction and the reward predictions of every other cue:

$$\begin{aligned} p(i) &= \frac{\exp(\beta \hat{m}^i)}{\sum_j \exp(\beta \hat{m}^j)} \\ &= \left(1 + \sum_{j \neq i} \exp(\beta (\hat{m}^j - \hat{m}^i)) \right)^{-1}, \end{aligned} \quad (9)$$

where β is a constant (analogous to the inverse temperature in thermodynamics) and modulates the extent to which $p(i)$ increases or decreases with respect to $\hat{m}^j - \hat{m}^i$. When $\beta = 0$, choices are independent of the learned valence, and each of the M available options are chosen with equal probability, $p(i) = M^{-1}$. When $\beta = \infty$, decisions are made deterministically, such that the cue with the highest predicted reward is always chosen. For the multiple-alternative forced choice task, the cue that is ultimately chosen on a given trial is determined by drawing a single, random sample, ξ , from a uniform distribution in the range 0-1, and selecting a cue, q , such that

$$q = \sup \left\{ \sum_{i=1}^x p(i) \leq \xi \mid 0 < x \leq M \right\}, x \text{ an integer.} \quad (10)$$

DAN firing rates

Once a cue has been chosen, the reward prediction specific to that cue is fed back to the DANs where they are compared against the actual reward, $\hat{r}^i = r_{+}^i - r_{-}^i$,

received in that trial, where $r_{\pm}$ is the magnitude of reward signal $R_{\pm}$. Given the chosen cue, q , $D_{\pm}$ firing rates in the VS models are given by

$$d_{\pm}^q = f(r_{\pm}^q + w_M m_{\mp}^q + \mathbf{w}_K^T \mathbf{k}^q), \quad (11)$$

whereas, in the MV model, $D_{\pm}$ is given by

$$d_{\pm}^q = f(r_{\pm}^q - r_{\mp}^q - w_M (m_{\mp}^q - m_{\pm}^q) + \mathbf{w}_K^T \mathbf{k}^q). \quad (12)$$

The difference in DAN firing rates yields the RPE for cue q :

$$\begin{aligned} \hat{d}^q &= d_+^q - d_-^q \\ &= \begin{cases} \hat{r}^q - \hat{m}^q, & \text{for VS models} \\ 2(\hat{r}^q - \hat{m}^q), & \text{for MV model} \end{cases} \end{aligned} \quad (13)$$

where $\hat{d}^q$ for the MV model is valid when $\mathbf{w}_K^T \mathbf{k}^q > |\hat{r}^q - \hat{m}^q|$. When the inequality is not satisfied, the precise expression for $\hat{d}^q$ in the MV model, taking into consideration the non-linear rectification in d_+ and d_- , is

$$\hat{d}^q = (\hat{r}^q - \hat{m}^q) + \text{sgn}(\hat{r}^q - \hat{m}^q) \min(\hat{r}^q - \hat{m}^q, \mathbf{w}_K^T \mathbf{k}^q). \quad (14)$$

Synaptic plasticity

We assume that the objective of the MB is to form accurate reward predictions, which minimise RPEs. This objective can be formulated as

$$\begin{aligned} C^{\text{RPE}} &= \frac{1}{2} \sum_i (\hat{r}^i - \hat{m}^i)^2 \\ &= \frac{1}{2} \sum_i (r_+^i - r_-^i - (\mathbf{w}_+^T \mathbf{k}^i - \mathbf{w}_-^T \mathbf{k}^i))^2, \end{aligned} \quad (15)$$

where the sum is over all cues, i , $\mathbf{w}_+^T \mathbf{k}^i$ is the firing rate of M_+ , expressed as the weighted input from the KC population response, $\mathbf{k}^i$, through synapses with strength $\mathbf{w}_+$, and similarly for $\mathbf{w}_-^T \mathbf{k}^i$. Learning an accurate reward prediction amounts to minimising C^{RPE} by modifying the synaptic weights. Assuming that inputs onto approach and avoidance MBONs are modified independently (18), we perform gradient descent on C^{RPE} with respect to $\mathbf{w}_+$ and $\mathbf{w}_-$ separately. The plasticity rule, $\mathcal{P}_{\pm}^{\text{RPE}}$, is then defined by the negative gradient:

$$\begin{aligned} \mathcal{P}_{\pm}^{\text{RPE}} &= -\eta \frac{\partial C^{\text{RPE}}}{\partial \mathbf{w}_{\pm}} \\ &= \eta \mathbf{k} (r_{\pm} - r_{\mp} - (\mathbf{w}_{\pm}^T \mathbf{k} - \mathbf{w}_{\mp}^T \mathbf{k})) \\ &= \frac{\eta}{2} \mathbf{k} (d_{\pm} - d_{\mp}), \end{aligned} \quad (16)$$

where η is the learning rate.

We take the same approach to derive a plasticity rule for the VS model. We start with the cost function

$$\begin{aligned} C_{\pm}^{\text{VS}} &= \frac{1}{2} \sum_i (r_{\mp}^i + m_{\pm}^i)^2 \\ &= \frac{1}{2} \sum_i (r_{\mp}^i + \mathbf{w}_{\pm}^T \mathbf{k}^i)^2, \end{aligned} \quad (17)$$

and compute the plasticity rule, $\mathcal{P}_{\pm}^{\text{VS}}$, by gradient descent:

$$\begin{aligned} \mathcal{P}_{\pm}^{\text{VS}} &= -\eta \frac{\partial C_{\pm}^{\text{VS}}}{\partial \mathbf{w}_{\pm}} \\ &= \eta \mathbf{k} (r_{\mp} + \mathbf{w}_{\pm}^T \mathbf{k}) \\ &= \eta \mathbf{k} (\mathbf{w}_{\pm}^T \mathbf{k} - d_{\mp}). \end{aligned} \quad (18)$$

The plasticity rule is in fact only an approximation to gradient descent, and holds true only when the learning rate, η , is sufficiently small. Here, sufficiently small means that $\eta < \left(2 \sum_j k_j^i\right)^{-1}$ for all cues, i , which ensures that learning does not result in unstable oscillations in reward predictions. In our simulations, we set $\eta = 2.5 \times 10^{-2}$. The plasticity rule therefore describes the mean drift in synaptic weights over several trials. In the simulations, we use Eqs. 16 and 18 to specify discrete updates to the synaptic weights at the end of each trial, t , conditioned on the chosen cue, q . Specifically, the update for the VS model is given by

$$\begin{aligned} \mathbf{w}_{\pm}(t+1) &= \mathbf{w}_{\pm}(t) + \mathcal{P}_{\pm}^{\text{VS}} \\ &= \mathbf{w}_{\pm}(t) + \eta \mathbf{k}^q(t) (\mathbf{w}_{\pm}^T(t) \mathbf{k}^q(t) - d_{\mp}^q(t)), \end{aligned} \quad (19)$$

and for the MV model by

$$\begin{aligned} \mathbf{w}_{\pm}(t+1) &= \mathbf{w}_{\pm}(t) + \mathcal{P}_{\pm}^{\text{RPE}} \\ &= \mathbf{w}_{\pm}(t) + \eta \mathbf{k}^q(t) (d_{\pm}^q(t) - d_{\mp}^q(t)). \end{aligned} \quad (20)$$

where the superscript q specifies the firing rate of each neuron in the presence of cue q alone, under the assumption that this cue dominates the neural activity at the point of receiving its corresponding reward signal. The update equation for the VS model with the modified plasticity rule (which we call the VS λ model) is

$$\begin{aligned} \mathbf{w}_{\pm}(t+1) &= \mathbf{w}_{\pm}(t) + \mathcal{P}_{\pm}^{\text{VS}\lambda} \\ &= \mathbf{w}_{\pm}(t) + \eta \mathbf{k}^q(t) (\lambda - d_{\mp}^q(t)). \end{aligned} \quad (21)$$

Note that the plasticity rule is not a function of the postsynaptic MBON firing rate (except indirectly through the DAN firing rate). This is possible because a separate plasticity rule exists for synapses impinging on each MBON, negating the need to label the postsynaptic neuron via its firing rate, as would be the case in three-factor Hebbian rules that are typically used in models of reward-modulated learning (26).

Reward schedule

At the end of each trial, a reinforcement signal specific to sensory cue i is provided. Rewards, r^i , take continuous values, and are drawn on each trial, t , from a normal distribution, $r^i(t) \sim \mathcal{N}(\mu_i(t), \sigma_R)$, with mean $\mu_i(t)$, and standard deviation σ_R . The reward signals that arrive at DANs, R_+ and R_- in Fig. 1D, have amplitudes $r_+^i = \max(0, r^i)$ and $r_-^i = \min(0, r^i)$, respectively. Over the course of a simulation run, $\mu_i(t)$ is varied according to a predetermined schedule, and σ_R is fixed. Thus, at different stages throughout each experiment, the most rewarding cue may switch between the multiple alternatives. Unless otherwise stated, $\sigma_R = 0.1$. The reward schedules were as follows. For Fig. 2 and 3, and Supplementary Fig. 1-4, $\mu_1(t=1) = 0$, and was held fixed for 20 trials, then underwent a step change of +1 at trials 21, 41, 141, and 161, and a step change of -1 at trials 61, 81, 101, and 121. For Fig. 4 and Supplementary Fig. 5, $\mu_i(t) = Ag(\xi_\mu(t)) + \sigma_R\xi_\sigma(t)$, where $\xi_\mu(t)$ and $\xi_\sigma(t)$ are Gaussian white noise processes with zero mean and unit variance, such that ξ_μ determines the mean reward, and $\xi_\sigma(t)$ determines the additive noise on trial t . A low pass filter, $g(\xi_\mu) = F^{-1}\{F\{\xi_\mu\}F\{G(0, \tau)\}\}$, is applied to ξ_μ , where $G(0, \tau)$ is a Gaussian function with unit area, centred on 0, and with standard deviation $\tau = 10$ trials, $F\{\cdot\}$ is the Fourier transform, and $F^{-1}\{\cdot\}$ is the inverse Fourier transform. Because the Fourier transform method of filtering assumes $\xi_\mu(1) = \xi_\mu(N_t + 1)$, where N_t is the number of trials, we generate ξ_μ for 250 trials, then delete the first 50 trials after filtering. Finally, the reward amplitude is determined by $A = 2 / \arg\max_t (g(\xi_\mu(t)))$.

Experimental data and model comparisons

The VS λ and MV models were compared to experimental data by simulating an often used conditioning protocol. To align with experiments, each simulation utilised the following procedure (Fig. 5a): *i*) in the first stage of training, the model is exposed to a single cue by itself, the CS+, for 10 trials, with rewards drawn from a normal distribution, $\mathcal{N}(\mu, 0.1)$, where μ was chosen according to whether appetitive ($\mu = 1$), aversive ($\mu = -1$), or neutral ($\mu = 0$) conditioning was simulated; *ii*) during the next 10 trials, the model is exposed to a second cue by itself, the CS-, with rewards drawn from a distribution with $\mu = 0$ and the same variance as for the CS+; *iii*) the final 2 trials comprise the test stage, in which the model is exposed to both cue 1 and cue 2, as in the MAFC task with two alternatives, with $\mu = 0$ for both cues. On each test trial, the model is forced to choose either cue 1 or cue 2, using Eq. 9. We used 10 trials per training stage as, given the parameters for η (learning rate) and β (inverse temperature), it took this many trials for the mean performance (see below for how performance is measured) across multiple runs of the simulation to plateau at, or near, the maximum possible value. The test was run for only 2 trials as synaptic plasticity was allowed to continue during the test stage, under the assumption that the formation of new CS+ related short term memories (13; 14) might alter the behaviour of flies in the

test stage of experiments.

For each simulation, we applied one of many possible additional protocol features, in which neuronal activity was manipulated. We therefore define a protocol as a unique combination of four features:

1. US valence (Fig. 5a): *i*) appetitive ($\mu = 1$); *ii*) aversive ($\mu = -1$); *iii*) neutral ($\mu = 0$).
2. Intervention type (Fig. 5c), which modified the target neuron's output firing rate from y_{targ} to $\tilde{y}_{\text{targ}}$: *i*) block of neuronal output (e.g. by *shibire*), which was simulated by multiplicatively scaling the manipulated neuron's firing rate, such that $\tilde{y}_{\text{targ}} = 0.1y_{\text{targ}}$; *ii*) neuronal activation (e.g. by dTrpA1), which was simulated by adding a constant current, such that $\tilde{y}_{\text{targ}} = y_{\text{targ}} + 5$.
3. The intervention type was applied following one of four intervention schedules (Fig. 5b): *i*) during the CS+ only; *ii*) throughout training (CS+ and CS-); *iii*) during test only; *iv*) throughout all stages.
4. The target neuron to which the intervention type was applied (Fig. 5c): *i*) M_+ ; *ii*) M_- ; *iii*) D_+ ; *iv*) or D_- .

We compared behavioural data from experiments with that of our model for 27 of the 96 possible variations of these four features. These data were obtained from a total of 439 experiments (235 controls with no intervention, 204 experiments with one of the 27 interventions), which were described in 14 published studies (7–13; 22; 23; 27; 30; 31; 33; 34).

Simulations were run in batches of 50, each batch yielding 100 choices from the two test trials. From these choices, we computed a performance index (PI_{mod}) given by

$$PI_{\text{mod}} = \left(\frac{n_+ - n_-}{n_+ + n_-} \right), \quad (22)$$

where n_+ is the number of choices for the CS+ and n_- for the CS-. A distribution of PIs for each protocol was obtained by running 20 such batches. The 439 PIs from the experimental data were extracted by eye from the 14 published papers. These PIs are computed in a similar way as for the model, but where n_+ and n_- correspond to the number of flies that approached the CS+ and CS-, respectively.

To measure the effect strength of each intervention in both the model and the experiments, we converted PIs into fractions of flies (or model runs) that chose the CS+, $f = (PI + 1) / 2$, then computed a test statistic, Δ_f , which compares f_c from control to f_i from intervention experiments, given that the underlying data is binomially distributed, as follows:

$$\Delta_f = \frac{f_i - f_c}{\sqrt{\frac{1}{N_{\text{fly}}} (f_i + f_c) \left(1 - \frac{1}{2} (f_i + f_c)\right)}}, \quad (23)$$

where N_{fly} is the number of flies used in that experiment. The binomial distribution adjustment to $f_i - f_c$ accounts for the bounded nature of f between 0 and 1. As such, for a given absolute difference, $f_i - f_c$, Δ_f is larger when f_c is near to 1 than when it is near to 0.5. That is, small changes to excellent memory performance imply a stronger effect than small changes to mediocre performance. Because N_{fly} was rarely stated in the studies we assessed, we set $N_{\text{fly}} = 50$, which is typical for experiments of this nature, and corresponds to the number of runs in each batch of simulations from which a single PI was computed from the model.

To examine the correspondence between PIs from the model and experiments, we fit a weighted linear regression to the experimental versus model Δ_f data using the MATLAB R2012a function *robustfit*, which computes iteratively reweighted least square fits with a bisquare weighting function. We then computed the Pearson correlation coefficient, R , of the weighted data using the weights, w_r , provided by *robustfit*, according to

$$R = \frac{\text{cov}(\mathbf{w}_r \Delta_f^{\text{mod}}, \mathbf{w}_r \Delta_f^{\text{exp}})}{\sigma_{\text{mod}} \sigma_{\text{exp}}}, \quad (24)$$

where σ_{mod} and σ_{exp} are the standard deviations of $\mathbf{w}_r \Delta_f^{\text{mod}}$ and $\mathbf{w}_r \Delta_f^{\text{exp}}$, respectively, and bold fonts denote vectors for all data points in either the model or experimental data sets. We determined the probability with which R comes from a distribution with zero mean by reshuffling the weighted data

Data availability

All experimental data was lifted from the data plotted in the cited publications. No additional experimental data was generated in this work. This data is collated and made available at:
https://github.com/BrainsOnBoard/MB_reinforcement_learning/mb_rescorla_wagner.

Code availability

All of the code that was used for running simulations and analysing data can be found on the Brains on Board GitHub repository:
https://github.com/BrainsOnBoard/MB_reinforcement_learning/mb_rescorla_wagner.

References

- [1] Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. & Prokasy, W. F. (eds.) *Class. Cond. II Curr. Res. theory*, 64–99 (Appleton-Century-Crofts, New York: Appleton-Century-Crofts, 1972).
- [2] Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–9 (1997). URL <http://www.ncbi.nlm.nih.gov/pubmed/9054347>.
- [3] Schultz, W. Neuronal Reward and Decision Signals: From Theories to Data. *Physiol. Rev.* **95**, 853–951 (2015). URL <http://www.ncbi.nlm.nih.gov/pubmed/26109341>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4491543>
<http://www.physiology.org/doi/10.1152/physrev.00023.2014>.
- [4] Waddell, S. Reinforcement signalling in *Drosophila*; dopamine does it all after all. *Curr. Opin. Neurobiol.* **23**, 324–329 (2013). URL <https://www.sciencedirect.com/science/article/pii/S0959438813000214>.
- [5] Schwaerzel, M. *et al.* Dopamine and octopamine differentiate between aversive and appetitive olfactory memories in *Drosophila*. *J. Neurosci.* **23**, 10495–10502 (2003). URL <http://www.ncbi.nlm.nih.gov/pubmed/14627633>.
- [6] Mao, Z. & Davis, R. L. Eight different types of dopaminergic neurons innervate the *Drosophila* mushroom body neuropil : anatomical and physiological heterogeneity MATERIALS AND METHODS. *Front. Neural Circuits* **3**, 1–17 (2009).
- [7] Burke, C. J. *et al.* Layered reward signalling through octopamine and dopamine in *Drosophila*. *Nature* **492**, 433–437 (2012).
- [8] Liu, C. *et al.* A subset of dopamine neurons signals reward for odour memory in *Drosophila*. *Nature* **488**, 512–516 (2012). URL <http://www.ncbi.nlm.nih.gov/pubmed/22810589>
<http://www.nature.com/articles/nature11304>.
- [9] Aso, Y. *et al.* Mushroom body output neurons encode valence and guide memory-based action selection in *Drosophila*. *Elife* **3** (2014). URL <https://elifesciences.org/articles/04580>.

- [10] Oswald, D. *et al.* Activity of Defined Mushroom Body Output Neurons Underlies Learned Olfactory Behavior in *Drosophila*. *Neuron* **86**, 417–427 (2015). URL <http://dx.doi.org/10.1016/j.neuron.2015.03.025>.
- [11] Perisse, E. *et al.* Aversive Learning and Appetitive Motivation Toggle Feed-Forward Inhibition in the *Drosophila* Mushroom Body. *Neuron* **90**, 1086–1099 (2016). URL <http://dx.doi.org/10.1016/j.neuron.2016.08.025>.
- [12] Ichinose, T. *et al.* Reward signal in a recurrent circuit drives appetitive long-term memory formation. *Elife* **4** (2015). URL <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4643015/>.
- [13] Felsenberg, J., Barnstedt, O., Cognigni, P., Lin, S. & Waddell, S. Re-evaluation of learned information in *Drosophila*. *Nature* **544**, 240–244 (2017). URL <http://dx.doi.org/10.1038/nature21716>.
- [14] Felsenberg, J. *et al.* Integration of Parallel Opposing Memories Underlies Memory Extinction. *Cell* **175**, 709–722.e15 (2018). URL <https://www.sciencedirect.com/science/article/pii/S0092867418310377>.
- [15] Zhao, X., Lenek, D., Dag, U., Dickson, B. J. & Keleman, K. Persistent activity in a recurrent circuit underlies courtship memory in *Drosophila*. *Elife* **7** (2018). URL <https://elifesciences.org/articles/31425>.
- [16] Aso, Y. *et al.* The neuronal architecture of the mushroom body provides a logic for associative learning. *Elife* **3**, e04577 (2014).
- [17] Cohn, R., Morante, I. & Ruta, V. Coordinated and Compartmentalized Neuromodulation Shapes Sensory Processing in *Drosophila*. *Cell* **163**, 1742–1755 (2015). URL <http://dx.doi.org/10.1016/j.cell.2015.11.019>.
- [18] Hige, T. *et al.* Heterosynaptic Plasticity Underlies Aversive Olfactory Learning in *Drosophila*. *Neuron* **88**, 985–998 (2015). URL <http://dx.doi.org/10.1016/j.neuron.2015.11.003>.
- [19] Bazhenov, M., Huerta, R. & Smith, B. H. A Computational Framework for Understanding Decision Making through Integration of Basic Learning Rules. *J. Neurosci.* **33**, 5686 (2013). URL <http://www.ncbi.nlm.nih.gov/pubmed/23536082> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3667960>.
- [20] Peng, F. & Chittka, L. A Simple Computational Model of the Bee Mushroom Body Can Explain Seemingly Complex Forms of Olfactory Learning and Memory. *Curr. Biol.* **27**, 224–230 (2017). URL <https://www.sciencedirect.com/science/article/pii/S096098221631288X>.
- [21] Cope, A. J. *et al.* Abstract concept learning in a simple neural network inspired by the insect brain. *PLOS Comput. Biol.* **14**, e1006435 (2018). URL <https://dx.plos.org/10.1371/journal.pcbi.1006435>.
- [22] Lin, S. *et al.* Neural correlates of water reward in thirsty *Drosophila*. *Nat. Neurosci.* **17**, 1536–1542 (2014).
- [23] Aso, Y. & Rubin, G. M. Dopaminergic neurons write and update memories with cell-type-specific rules. *Elife* **5** (2016).
- [24] Bouzaiane, E., Trannoy, S., Scheunemann, L., Plaçais, P.-Y. & Preat, T. Two Independent Mushroom Body Output Circuits Retrieve the Six Discrete Components of *Drosophila* Aversive Memory. *Cell Rep.* **11**, 1280–1292 (2015). URL <https://www.sciencedirect.com/science/article/pii/S2211124715004611>.
- [25] Séjourné, J. *et al.* Mushroom body efferent neurons responsible for aversive olfactory memory retrieval in *Drosophila*. *Nat. Neurosci.* **14**, 903–910 (2011). URL <http://www.nature.com/articles/nn.2846>.

- [26] Frémaux, N. & Gerstner, W. Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. *Front. Neural Circuits* **9**, 85 (2015). URL <http://www.ncbi.nlm.nih.gov/pubmed/26834568> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4717313>.
- [27] Huetteroth, W. *et al.* Sweet taste and nutrient value subdivide rewarding dopaminergic neurons in drosophila. *Curr. Biol.* **25**, 751–758 (2015).
- [28] Yamagata, N. *et al.* Distinct dopamine neurons mediate reward signals for short- and long-term memories. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 578–83 (2015). URL <http://www.ncbi.nlm.nih.gov/pubmed/25548178> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4299218>.
- [29] Cervantes-Sandoval, I., Phan, A., Chakraborty, M. & Davis, R. L. Reciprocal synapses between mushroom body and dopamine neurons form a positive feedback loop required for learning. *Elife* **6:e23789** (2017).
- [30] Yamagata, N., Hiroi, M., Kondo, S., Abe, A. & Tanimoto, H. Suppression of Dopamine Neurons Mediates Reward. *PLOS Biol.* **14**, e1002586 (2016). URL <http://www.ncbi.nlm.nih.gov/pubmed/27997541> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5172549> <http://dx.plos.org/10.1371/journal.pbio.1002586>.
- [31] Plačais, P.-Y., Trannoy, S., Friedrich, A. B., Tanimoto, H. & Preat, T. Two Pairs of Mushroom Body Efferent Neurons Are Required for Appetitive Long-Term Memory Retrieval in Drosophila. *Cell Rep.* **5**, 769–780 (2013). URL <http://www.ncbi.nlm.nih.gov/pubmed/24209748> <https://linkinghub.elsevier.com/retrieve/pii/S221112471300555X>.
- [32] Riemensperger, T., Völler, T., Stock, P., Buchner, E. & Fiala, A. Punishment Prediction by Dopaminergic Neurons in Drosophila. *Curr. Biol.* **15**, 1953–1960 (2005). URL <http://www.ncbi.nlm.nih.gov/pubmed/16271874> <https://linkinghub.elsevier.com/retrieve/pii/S0960982205011127>.
- [33] Claridge-Chang, A. *et al.* Writing Memories with Light-Addressable Reinforcement Circuitry. *Cell* **139**, 405–415 (2009). URL <https://www.sciencedirect.com/science/article/pii/S0092867409011040>.
- [34] Aso, Y. *et al.* Specific Dopaminergic Neurons for the Formation of Labile Aversive Memory. *Curr. Biol.* **20**, 1445–1451 (2010). URL <https://www.sciencedirect.com/science/article/pii/S0960982210007967>.
- [35] Möller, M. & Bogacz, R. Learning the payoffs and costs of actions. *PLOS Comput. Biol.* **15**, e1006285 (2019). URL <http://dx.plos.org/10.1371/journal.pcbi.1006285>.
- [36] Dylla, K. V., Raiser, G., Galizia, C. G. & Szyszka, P. Trace Conditioning in Drosophila Induces Associative Plasticity in Mushroom Body Kenyon Cells and Dopaminergic Neurons. *Front. Neural Circuits* **11**, 42 (2017). URL <http://journal.frontiersin.org/article/10.3389/fncir.2017.00042/full>.
- [37] 1988sutton.s, R. S. Learning to Predict by the Methods of Temporal Differences. *Mach. Learn.* **3**, 9–44 (1988). URL <https://link.springer.com/content/pdf/10.1007%2FBF00115009.pdf>.
- [38] 1988sutton.s, R. S. & Barto, A. G. *Reinforcement learning : an introduction* (The MIT Press, Cambridge, MA, 2018), 2nd edn. URL <https://mitpress.mit.edu/books/reinforcement-learning-second-edition>.
- [39] Takemura, S.-Y. *et al.* A connectome of a learning and memory center in the adult Drosophila brain. *Elife* **6** (2017). URL <http://www.ncbi.nlm.nih.gov/pubmed/28718765> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5550281>.

- [40] Eichler, K. *et al.* The complete connectome of a learning and memory centre in an insect brain. *Nature* **548**, 175–182 (2017). URL <http://www.nature.com/articles/nature23455>
<http://www.ncbi.nlm.nih.gov/pubmed/28796202> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4000970>
- [41] Tanimoto, H., Heisenberg, M. & Gerber, B. Event timing turns punishment to reward. *Nature* **430**, 983–983 (2004). URL <http://www.nature.com/articles/430983a>.
- [42] Klopff, A. H. Brain Function and Adaptive Systems-A Heterostatic Theory. Technical Report AFCRL- 72-0164. Tech. Rep., Air Force Cambridge Research Laboratories, Bedford, MA (1972). URL <http://www.dtic.mil/dtic/tr/fulltext/u2/742259.pdf>.
- [43] Cognigni, P., Felsenberg, J. & Waddell, S. Do the right thing: neural network mechanisms of memory formation, expression and update in *Drosophila*. *Curr. Opin. Neurobiol.* **49**, 51–58 (2017). URL <http://dx.doi.org/10.1016/j.conb.2017.12.002>.
- [44] Robbins, H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **58**, 527–536 (1952). URL <http://www.ams.org/journal-getitem?pii=S0002-9904-1952-09620-8>.
- [45] Wang, Y. *et al.* Stereotyped Odor-Evoked Activity in the Mushroom Body of *Drosophila* Revealed by Green Fluorescent Protein-Based Ca²⁺ Imaging. *J. Neurosci.* **24**, 6507–6514 (2004).
- [46] Turner, G. C., Bazhenov, M. & Laurent, G. Olfactory Representations by *Drosophila* Mushroom Body Neurons. *J. Neurophysiol.* **99**, 734–746 (2008).
- [47] Lin, A. C., Bygrave, A. M., de Calignon, A., Lee, T. & Miesenböck, G. Sparse, decorrelated odor coding in the mushroom body enhances learned odor discrimination. *Nat. Neurosci.* **17**, 559–68 (2014). URL <http://www.ncbi.nlm.nih.gov/pubmed/24561998>
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4000970>.
- [48] Technau, G. & Heisenberg, M. Neural reorganization during metamorphosis of the corpora pedunculata in *Drosophila melanogaster*. *Nature* **295**, 405–407 (1982). URL <http://www.nature.com/articles/295405a0>.

Acknowledgements

Special thanks to Eleni Vasilaki for helpful discussions and feedback on the mathematical formulations, James Marshal for feedback on the manuscript, and the Waddell and Vogels labs for fruitful discussions on learning in *Drosophila*. Thanks also to the members of the Brains on Board team for their critical feedback at various points throughout this project. This work was funded by the EPSRC (Brains on Board project, grant number EP/P006094/1).

Author contributions

J.E.M.B. conceived the model, wrote the code, generated and analysed the data. J.E.M.B. and T.N. conceived the reward schedules and ideal agents to test the models. J.E.M.B., A.P. and T.N. wrote and revised the manuscript.

Figures

Figure 1: Valence-specific model of the mushroom body. **a** 3D render of several neuropils that comprise the brain of *Drosophila melanogaster*. The red region highlights the MB in the right hemisphere. Labels: MB – mushroom body; AL – antennal lobe; SOG – suboesophageal ganglion; ME – medulla. **b** Outlines of the multiple compartments that tile the lobes of the MB, colour-coded by a broad classification of cell function. Blue: approach MBONs; orange: avoidance MBONs; purple: aversive DANs; green: appetitive DANs; black: KCs. *Inset*: schematic of the three MB lobes and their compartmentalisation. Top: α'/β' lobes; middle: α/β lobes; bottom: γ lobe. MBON functions (approach or avoidance) are as determined in (9). Pale colours in the *inset* correspond to MBONs that exhibit a non-significant bias on behaviour in (9). **c** Schematic of the VS model. Units are colour-coded according to the cell types in **b**. KCs connect to MBONs through plasticity synapses, and connect to DANs through fixed synapses. Labels: M_+ – approach MBON; M_- – avoidance MBON; D_- – aversive DAN; D_+ – appetitive DAN; K – Kenyon cells; R_- – negative reward; R_+ – positive reward. Lines with arrows: excitatory synapse. Lines with filled circles: synapse releasing dopamine. Downward white arrows: dopamine enhances synaptic depression. **d** Schematic of a single trial for the experimental paradigm in which the model is examined. Panel 1: the model is exposed to some number of cues that are evaluated to yield a cue-specific reward prediction. Panel 2: using the relative reward predictions, the model makes a decision over which cue to choose. Panel 3: the model receives a reinforcement signal that is associated with the chosen cue, and its reward predictions for that cue are updated.

Figure 2: Reward predictions in the valence-specific model track rewards but only within specified bounds. **a** Reward schedule (excluding the Gaussian white noise; thick, light blue) and reward predictions (thin, various dark blues) from 10 separate runs of the model. Each shade of dark blue corresponds to simulations using a specific KC→DAN synaptic weight, which is determined by γ : dark blue through to very dark blue corresponds to $\gamma = 0.9, 1.0, 1.1$, and $\gamma > 1.15$. Dashed lines correspond to $\hat{m}_{\max}$, the theoretical maximum to the reward prediction magnitude, for each value of γ . **b** Firing rates of the M_+ (blue) and the M_- MBONs (orange) for 10 runs of the model, with the same reward schedule as in **a**. **c** Firing rates for the D_+ (green) and the D_- (purple) DANs in response to the reward schedule in **a**. **d** RPEs given by the difference in firing rates of D_+ and D_- .

Figure 3: Dual versions of unbounded valence-specific (VSu) models can be combined to create the mixed-valence (MV) model, which also learns unbounded reward predictions. **a** One of the dual, VSu models that requires D_- and D_+ to be inhibited by positive and negative rewards, respectively. Lines with flat ends correspond to inhibitory synapses. **b** The second of the dual VSu models, in which MBONs provide inhibitory feedback to DANs of the same valence. Right, upward arrows in the dopamine synapses denote that dopamine induces synaptic potentiation. **c** The MV model, which combines the dual VSu models. Grey units are inhibitory interneurons. **d-e** Each panel exhibits the behaviour from 10 independent runs of the model. **d** Reward predictions are unbounded and accurately track the rewards, but the learning speed depends on KC→DAN synaptic weights. Thick, light blue: reward schedule; thin, dark blue: $\gamma = 0$; thin, very dark blue: $\gamma \geq 0.3$. *Inset*: magnified view of the region highlighted by the dashed square, showing how learning is slower when $\gamma = 0$. **e** M_+ (blue) and M_- firing rates when $\gamma = 1$. **f** D_+ and D_- firing rates when $\gamma = 1$. **g** RPEs as given by the difference between D_+ and D_- firing rates when $\gamma = 1$.

Figure 4: Learning reward predictions in tasks with multiple cues. **a** Reward schedules (lines) and reward predictions (circles, shown only for the cue chosen on each trial) for two cues (blue: cue 1; yellow: cue 2). Reward predictions are shown for 10 independent runs of a simulation using the same reward schedule. **b** Reward schedules (lines) and reward predictions (circles, shown only for the cue chosen on each trial) for a single run of the model in a task involving 5 cues. Each colour corresponds to rewards and predictions for a different cue.

Figure 5: Both the modified valence-specific (VS λ) and mixed valence (MV) models produce choice behaviour that corresponds well with experiments under a broad range of experimental manipulations. **a**) Schematic of the experimental protocol used to simulate appetitive, aversive and neutral conditioning experiments. **b-c**) The protocol was extended to simulate genetic interventions used in experiments. **b**) Interventions were applied at different stages of a simulation, either 1) during CS+ exposure in training, 2) during CS+ and CS- exposure in training, 3) during testing, and 4) throughout both training and testing. **c**) Four examples of the interventions simulated, each corresponding to encircled data in **d** and **e**. Red crosses denote interventions that simulate activation of a *shibire* blockade; yellow stars denote interventions that simulate activation of an excitatory current through the dTrpA1 channel. The picture at the top of each panel denotes the reward type, and the encircled number the activation schedule as specified in **b**. **d**) Comparison of PIs measured from the VS λ model and from experiments. **e**) Comparison of PIs measured from the MV model and from experiments. Grey lines in **d** and **e** are weighted least square linear fits with correlation coefficients $R = 0.65$ and $R = 0.63$ respectively. The size of each data point scales with its weight in the linear fit.