Dopamine reports reward prediction errors, but does not update policy, during inference-guided choice

Marta Blanco-Pozo[1#], Thomas Akam[1#], Mark Walton[1,2*]

Affiliations:

1. Department of Experimental Psychology, Oxford University, Oxford, UK

2. Wellcome Centre for Integrative Neuroimaging, Oxford University, Oxford, U.K.

# Equal contribution

mark.walton@psy.ox.ac.uk

thomas.akam@psy.ox.ac.uk

marta.blancopozo@psy.ox.ac.uk

Abstract

Dopamine is thought to carry reward prediction errors (RPEs), which update values and hence modify future behaviour. However, updating values is not always the most efficient way of adapting to change. If previously encountered situations will be revisited in future, inferring that the *state* of the world has changed allows prior experience to be reused when situations are reencountered. To probe dopamine's involvement in such inference-based behavioural flexibility, we measured and manipulated dopamine while mice solved a sequential decision task using state inference. Dopamine was strongly influenced by the value of states and actions, consistent with RPE signalling, using value information that respected task structure. However, though dopamine responded strongly to rewards, stimulating dopamine at the time of trial outcome had no effect on subsequent choice. Therefore, when inference guides choice, rewards have a dopamine-independent influence on policy through the information they carry about the world's state.

Introduction

Accurately estimating future reward associated with different actions is critical to adaptive behaviour. Reinforcement learning (RL) has provided an influential mechanistic account in which reward prediction errors (RPEs) update values (i.e., predictions of long run reward). Dopamine activity consistent with RPEs is observed in diverse behaviours[1–4], and causal manipulations of dopamine can reinforce or suppress behaviours[5–9].

However, RPE-driven value learning is not always the most efficient way of updating preferences in dynamic environments. In many situations, changes in observable events reflect underlying dynamics which are themselves highly structured. For example, the cyclical change between week and weekend constitutes a *latent* or not directly observable feature of our environment, which affects observable events, with consequences for behaviour. It would be profoundly inefficient to relearn values each Monday and Saturday to cope with this change; instead, by *inferring* that the *state* of the world has changed, we can just redeploy values and choice policies from a previous encounter. Such structured change is also common in the laboratory, and state inference could in principle underlie behavioural adaptation in many reward-guided tasks[10,11]. Indeed, direct evidence for inference based behavioural flexibility has been demonstrated in several tasks[12–16], and is also suggested by meta-learning, where subjects get faster at adapting to changes with repeated experience[17–21].

Therefore, a pressing issue is to understand how dopamine reflects and shapes behaviour during inference-guided choice. Dopamine activity is informed by inferences about latent states[22–27]. However, little is known about activity in more complex tasks where latent states and causal relationships interact during sequential decision making. This is ubiquitous in the real world; for instance, in the earlier example, when the best route home depends on the day of the week. We also do not know how the observed heterogeneity in dopamine projections to different targets represents such behaviours[6,28–33]. Most critically, in situations where rewards provide information about the *state* of the world, we do not know if their effect on subsequent behaviour is mediated by the RPEs they evoke (putatively carried by dopamine), or through influences on subject's beliefs.

To address these questions, we measured and manipulated dopamine in mice performing a two-step decision task. The task had two important features. First, reward probabilities were

anti-correlated and reversed periodically, thus introducing an inferable latent variable. Second, the action-state transition structure meant that under an optimal policy, rewards did not simply reinforce the preceding choice (Figure 1). Both behaviour and dopamine activity showed evidence that mice inferred a single latent variable that controlled both reward probabilities, and assigned value to actions consistent with the task's structure. Dopamine activity was strongly modulated at specific time points across the trial by the values of their chosen actions, the states they encountered, and the outcome they obtained. This coexisted with a separate sustained influence of recent reward rate throughout the trial, and signals correlated with lateralised movements. Nonetheless, stimulating midbrain dopamine neurons at outcome time, using parameters sufficient to reinforce preceding actions, did not recapitulate the effect of natural rewards, and had no effect on subsequent choice. These data demonstrate that when state inference underlies flexible behaviour, dopamine has a limited causal influence on choices, despite prominently reporting a set of task-related variables.

Results

Behaviour reflects task causal and correlation structure

We trained DAT-cre mice (n=18) on a sequential decision task, which required them to choose between two options – left and right – to gain access to one of two reward ports – top or bottom (Figure 1A). Each trial started with a central port lighting up, which mice poked to initiate the trial. This triggered either a choice state, where both left and right ports lit up (75% of trials), or a forced choice where either left or right lit up (25% of trials). Poking an illuminated side port triggered a transition to one of two possible second-step states, signalled by an auditory cue, in which either the top or bottom port was illuminated. Each first-step choice (left or right) led commonly (80% of trials) to one second-step state (up or down) and rarely (20% of trials) to the other (Figure 1B). Poking the active reward port caused a 500-ms auditory cue indicating the trial outcome (reward or not), with reward delivered at cue offset. Reward delivery in each port was probabilistic, with probabilities that changed in blocks between *0.8/0.2*, *0.5/0.5* and *0.2/0.8* on the up/down ports respectively. The reward probabilities were therefore anticorrelated at the up and down ports. Though we have previously used a two-step task where both transition and reward probabilities changed over

time[34], here the transition probabilities were fixed and only reward probabilities changed. This facilitates latent-state inference strategies because the latent state of the reward probabilities fully determines the first-step action values and optimal policy[11,34].

Subjects tracked which option was currently best, performing 348.70 ± 93.90 (mean ± SD across subjects) trials and completing 8.32 ± 3.24 blocks per session (Figure 1C). Rewards did not simply reinforce the choice made at the start of the trial, but rather the trial outcome (rewarded or not) interacted with the state transition (common or rare) to determine the probability of repeating choice (Figure 1D-F). Rewards following common transitions strongly promoted staying while rewards following rare transitions strongly promoted switching ($P < 0.001$, mixed effects logistic regression) (Figure 1D), and these effects persisted over multiple trials (Figure 1F).

This interaction between outcome and transition is consistent with model-based RL – where subjects evaluate first-step actions by combining the transition probabilities linking them to second-step states with the values of those states[29]. However, it is also consistent with a different type of strategy based on latent state inference[11]. A subject who can infer the latent state of the reward probabilities – i.e., whether reward probability is currently high at the up or down port – can in principle learn separate cached values for available actions and observable trial events conditioned on this. Cached values learnt under a latent state with high reward probability at the up port will be high for the 'up-active' second-step state and for the first-step action that commonly leads to this state, and low for the 'down-active' state and the alternative first-step choice. These values will be reversed for a latent state with high reward probability at the down port. Under this account, rewards shape subjects' *beliefs* about the current latent state, determining which set of values is in play, and hence the preferred first-step action. Subjects can then adapt to reversals by changing their beliefs about the current state, without needing to relearn values.

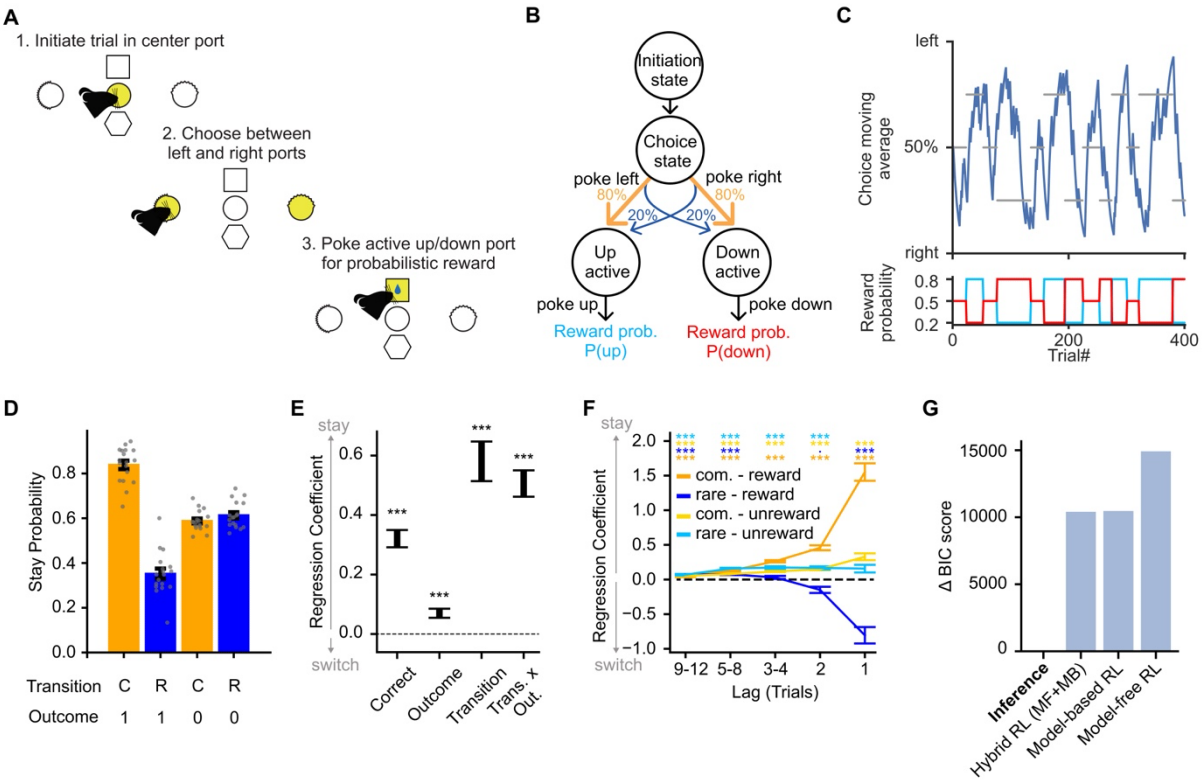


Figure 1. Two step task behaviour. A) Diagram of trial events. **B)** Diagram of the task state space, reward and transition probabilities. **C)** Example behavioural session. Top panel: Moving average of choices (blue trace) with reward probability blocks shown by grey bars, indicating by their vertical position whether the reward probability was higher for the state commonly reached for the left or right choice or a neutral block. Bottom panel: Reward probability at the up (blue) and down (red) ports. **D)** Probability of a choice being repeated as a function of the subsequent state transition (common or rare) and trial outcome (rewarded or unrewarded). Dots show individual subjects, error bars cross subject SEM. **E)** Mixed effects logistic regression predicting repeating choice as a function of the trial outcome (rewarded or not), state transition (common or rare), the transition-outcome interaction, and whether the choice was to the correct (high reward probability) option. **F)** Lagged logistic predicting choice as a function of the history of different trial types defined by the transition and outcome. **G)** Model comparison showing difference in BIC score for different models relative to the best model. Only the best model from each class of models considered is shown, the full set of models compared is shown in Figure S1A. *p<0.05, **p<0.01, ***p<0.001

Model comparison supported an inference-based account over standard model-based RL (Figure 1G, S1A). The best fitting model tracked the current state of the reward probabilities using a single latent variable updated in opposite directions by reward at the up and down ports, which in turn determined the relative value of the two first-step actions – i.e., if the model had recently received more reward at the up port, it favoured the first-step action which commonly led to that port. It therefore did not explicitly infer which discrete state the reward probabilities were in, but rather approximated this by a continuous latent variable which captured their anticorrelated structure. In addition to this 'inference' component, it incorporated a simpler strategy termed 'Reward-as-cue' (see ref [11]), which used the most

recent outcome (e.g., reward at the up-port) as a discriminative stimulus for which first-step action to choose on the next trial. Both components therefore used the trial outcome as information about the task's state, but differed with respect to whether they maintained this information beyond the subsequent choice.

Unexpectedly, both the stay probability and lagged regression analysis (Figure 1D,F) indicated a strong asymmetry in the effect of reward and reward omission on future choice. The state transition – and hence *where* the outcome was obtained – strongly affected future choices on rewarded trials but only had marginal effects on non-rewarded trials. This was confirmed by model comparison; reward at the up and down ports had strong opposing influences on the best fitting model's choice probability, updating the latent variable used by the inference component in opposite directions, and on the next trial transiently promoting the first-step action that commonly led to where the reward was obtained. By contrast, reward omission neither updated the latent variable nor transiently promoted one or other next-trial choice. This model fit the data better than standard RL models which learn values independently for the two second-step states, irrespective of whether these were model-free, model-based or a mixture of both, including models with asymmetric learning rates for positive and negative outcomes (Supplementary Tables 1 & 2). Behaviour simulated from the best model closely recapitulated subjects' data (Figure S1B).

Value information in dopamine activity respects task causal and correlation structure

To assess how task structure influenced dopamine activity, we used fibre photometry to record calcium signals from GCaMP6f expressing dopamine neuron cell bodies in the ventral tegmental area (VTA) and axons in nucleus accumbens (NAc) and dorsomedial striatum (DMS), and dopamine signals in NAc and DMS using dLight1.1 expressed pan-neuronally in striatum (see Figure S2 for placements).

Dopamine activity fluctuated dynamically across the trial, as the mice made their initial choice, received information about the second-step state reached, and discovered the trial outcome (Figure 2). Reward responses were prominent in all signals, though relatively weaker in DMS calcium. However, average DMS calcium activity masked a strong medio-lateral gradient in reward responsivity, with larger reward responses more laterally in DMS (Figure S3). For the following analyses, we therefore included only DMS sites > 1.2mm from

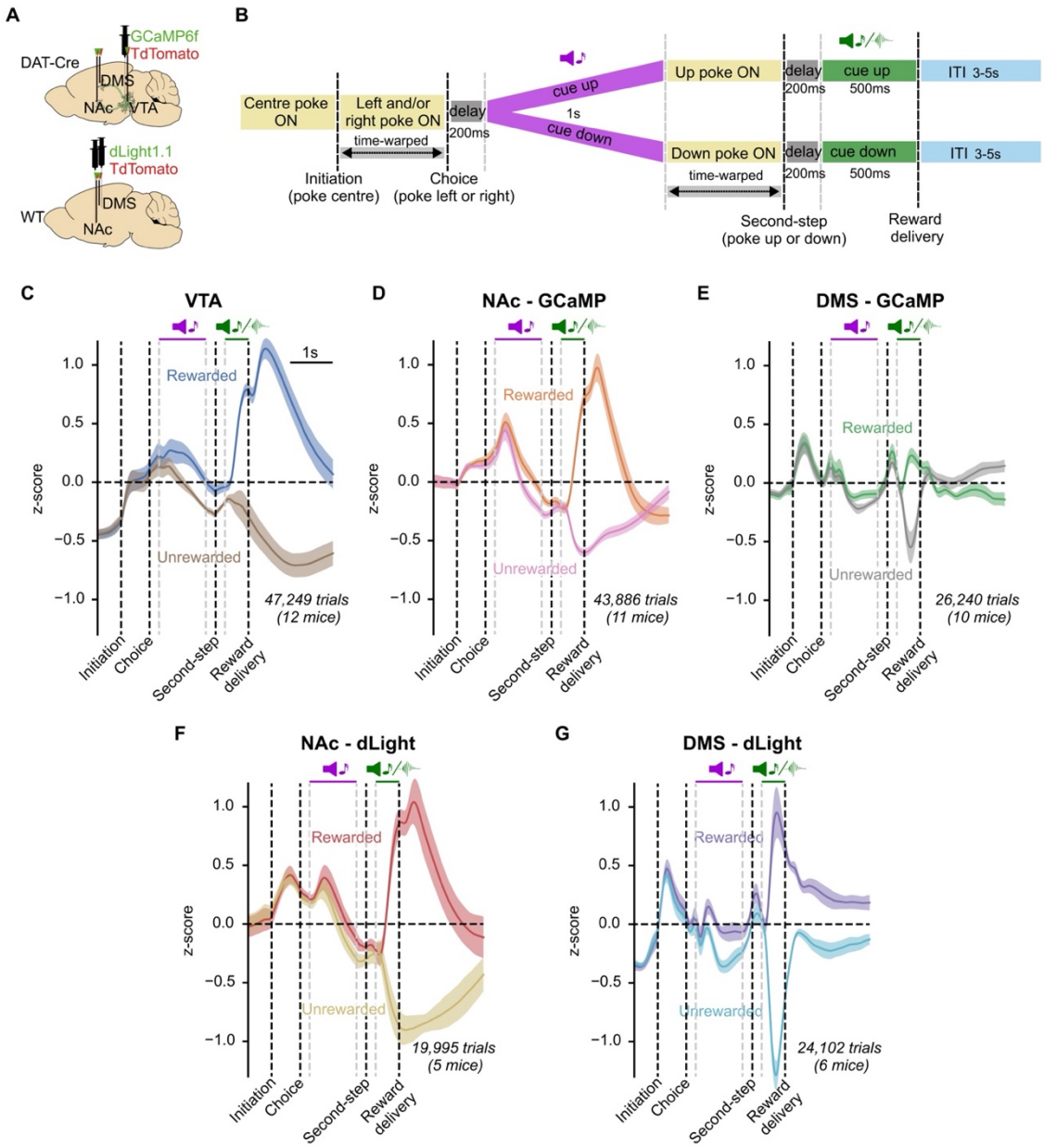


Figure 2. Dopamine activity and release are modulated across the trial. A) Injection and implant schematic. For calcium photometry, DAT-Cre mice were injected bilaterally with GCaMP6f and TdTomato in VTA, and optic fibres were implanted in VTA and NAc in one hemisphere, and DMS in the other hemisphere. For dLight photometry, WT mice were injected with dLight1.1 and TdTomato in NAc and DMS in different hemispheres, and optic fibres implanted at the same sites. **B)** Trial timeline. Trials were aligned by warping the intervals between the trial initiation and choice, and between the second-step port lighting up and being poked. **C-E)** Mean z-scored calcium signal on rewarded and unrewarded trials with shaded areas indicating cross-subject standard error in **C)** VTA (n: 12 mice, 121 sessions, 47,249 trials), **D)** NAc (n: 11 mice, 108 sessions, 43,886 trials), **E)** DMS (n: 10 mice, 74 sessions, 26,240 trials), **F-G)** As C-E but for NAc dopamine concentrations in **F)** NAc (n: 5 mice, 64 sessions, 19,995 trials), and **G)** DMS (n: 6 mice, 91 sessions, 24,102 trials.

bregma, excluding 2 animals where the fibre was most medial and where we observed a negative response to reward.

To resolve other influences on dopamine activity, we used linear regression to model dopamine fluctuations across trials and time-points. We aligned activity across trials (Figure 2B) and ran a separate regression analysis for each timepoint in the aligned data, using predictors which varied from trial-to-trial but took the same value for all timepoints in each trial. The time courses of predictor loadings across the trial therefore reflect when each predictor explained variance in the activity. We focus on the GCaMP data in the main figures, but results from dLight were closely comparable, except where noted in the text (see Figure S5-7).

Our first analysis was designed to test specific hypotheses about how the previous trial's events influenced value estimates for actions and states experienced on the current trial. These hypotheses were expressed as a set of regressors detailed below, but the analysis also included regressors modelling trial outcome and other possible sources of variance (see Methods, Figure S4,5). As expected, dopamine activity in all regions was strongly positively modulated by trial outcome (rewarded or not) from when this information became available (Figure 3A). However, while all regions responded to the (fully predictive) outcome cue, responses to subsequent reward delivery were strong in VTA, weaker in NAc and absent in DMS.

We next looked for evidence of subjects inferring a single latent variable that controlled both reward probabilities, using two regressors that coded the previous trial's outcome (rewarded or not) either on trials where the second-step state was (i) the same as that reached on the previous trial, or (ii) different from the previous trial. When the second-step state was the same, reward on the previous trial *increased* dopamine at the time the second-step was revealed on the current trial (Figure 3B). This is predicted by both independent value learning at the two second-step states and by inference about a single latent variable that controls both reward probabilities. However, when the second-step state was *different* from the previous trial, reward on the previous trial had a *negative* influence on dopamine at the time the second-step state was revealed on the current trial. This is not predicted by independent value learning – because reward at one second-step state does not affect the value of the other second-step state – but *is* predicted by inference of a single latent variable: evidence one state is good is also evidence the other is bad. In NAc, but, unexpectedly, in neither other region, loading on these regressors reversed sign at outcome time. This is exactly as predicted

for an RPE signal: at a state transition, the value of the new state contributes positively to the RPE, while the value of the old state contributes negatively[35–37].

We next assessed value information informing the dopamine response to subject's first-step choice, using two different regressors. A *direct-reinforcement* regressor coded the previous trial outcome (rewarded or not) only on trials where the current choice was the same as the previous choice (Figure 3C) – i.e., the direction in which model-free direct reinforcement of the previous choice by trial outcome would update the value of the first-step action chosen on the current trial. This regressor loaded positively around choice time in VTA and NAc, but not DMS, consistent with action value estimates used by mesolimbic dopamine incorporating a simple model-free update over observable events.

A second *model-consistent* predictor coded the direction in which the previous trial's outcome would affect the value of the action chosen on the current trial, given a correct model of both the anticorrelated reward probabilities and transition structure linking first-step actions to second-step states (Figure 3D). This loaded positively around choice time in VTA, NAc and DMS calcium activity, consistent with the dopamine system having access to value information about the chosen action that respected the task structure. Both the second-step value updates and model-consistent first-step action value update were primarily driven by rewarded trials and not reward omissions (Figure S6), consistent with the asymmetric effect of reward and omission on behaviour.

In addition to these value-related signals, dopamine was also modulated by lateralised movement (Figure S4, 5). Consistent with previous reports[6], activity in DMS but not VTA or NAc increased at the time of contralateral choices. However, each trial also required a second lateralised movement from choice port back to the centrally located second-step port, which did not evoke a response in DMS. Unexpectedly, VTA and NAc showed a lateralized response during this second contralateral movement from choice to second-step ports, despite not responding to the initial choice.

These data demonstrate that dopamine signals were shaped by value information that respected two aspects of task structure: the anticorrelated reward probabilities and the action-state transitions. Mesolimbic dopamine also carried a signature of simpler model-free value updating of actions by subsequent outcomes.

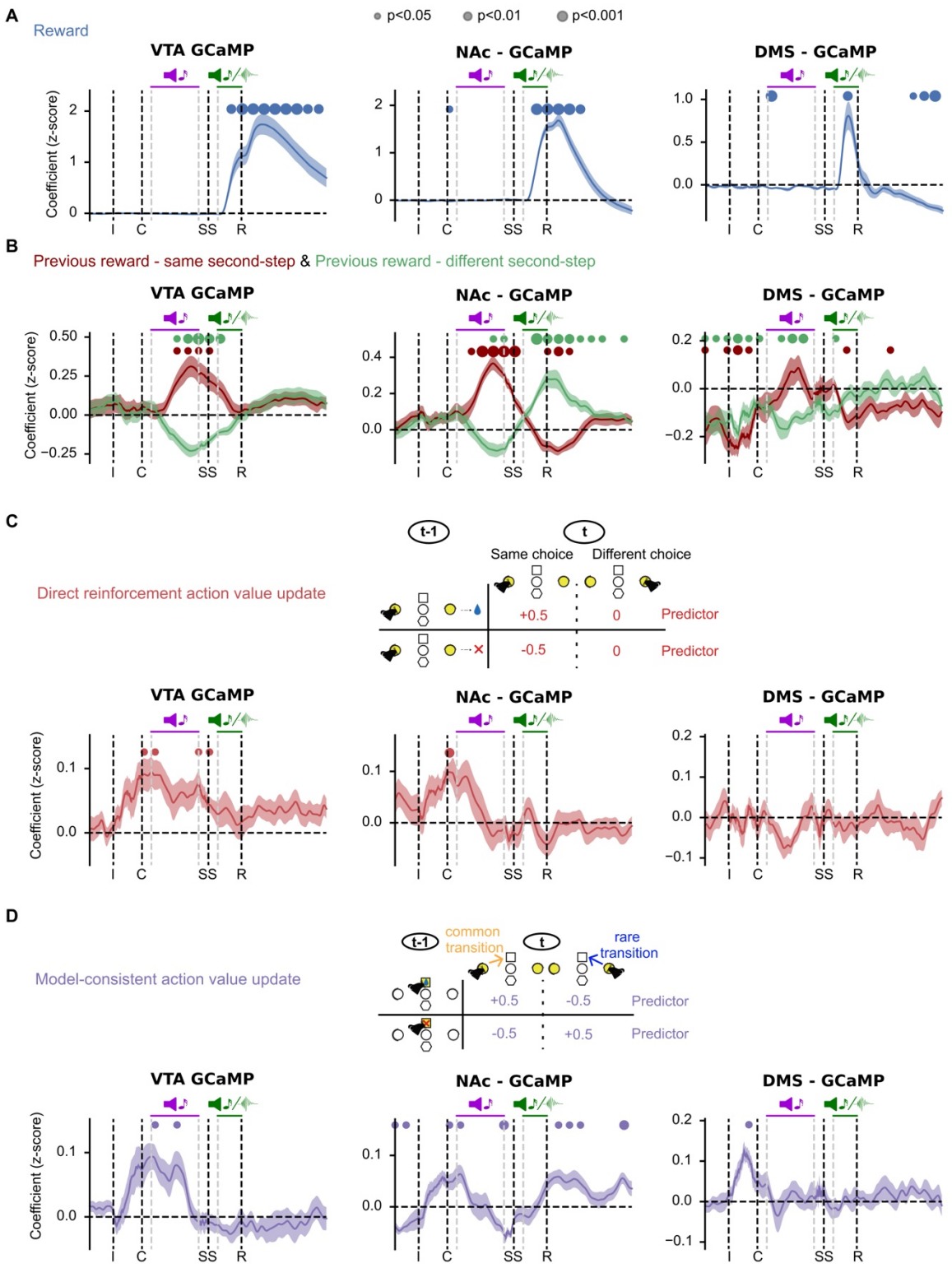


Figure 3. Dopaminergic value information respects task structure A-D) Coefficients in linear regression predicting dopamine activity for each timepoint in the trial for each region, showing cross subject-mean and standard error. Dots indicate timepoints where coefficients are statistically significant, assessed by a two-sided t-test comparing the cross-subject distribution against 0, Benjamini-Hochberg corrected for comparison of multiple time-points, small dot: $p < 0.05$, medium dot: $p < 0.01$, large dot: $p < 0.001$. **A)** Current trial outcome. **B)** Previous trial outcome when the second-step state was the same (red) or different (green) from the previous trial. **C)** Direct reinforcement action value update. Regressor codes the direction which the value

of the first-step action chosen on the current trial would be updated by direct reinforcement of the previous trial's first-step choice by the previous trial's outcome. Regressor takes values *+0.5/-0.5* when the current trial choice is the *same* as the previous trial and the previous outcome was *rewarded/not-rewarded*, *0* if current choice is *different* from previous choice. **D**) Model-consistent action value update. Regressor codes direction in which the value of the first-step action chosen on the current trial would be updated by the previous trial's outcome given knowledge of the task's causal and correlation structure. Regressor takes values *+0.5/-0.5* when the previous outcome is *rewarded/not-rewarded* and the current trial's first step choice *commonly* leads to the previous trial's second-step state, and *-0.5/+0.5* when the previous outcome is *rewarded/not-rewarded* and the current trial's first step choice *rarely* leads to the previous trial's second-step state.

Dissociable influence of value and reward rate on dopamine activity

The above analysis used predictors expressing specific hypotheses about how the previous trial's events affected value information in the dopamine system. A complementary picture can be obtained using trial-by-trial estimates of decision variables from the best fitting behavioural model to predict dopamine fluctuations. We used the same linear regression framework but with behavioural model values for the first-step choice and second-step state as predictors. As detailed above, the best fitting model used a single latent variable to track which option is best, updated in opposite directions by rewards at the two second-step states. The second-step value is just this variable interacted with the second-step state reached on the trial, while the chosen action value is this variable interacted with the chosen action. Both values therefore respect the anticorrelated reward probabilities, and the chosen action value combines this with the action-state transition structure.

Second-step state value had a strong positive influence on VTA and NAc calcium at the time the second-step was revealed, which subsequently reversed at outcome time, consistent with RPE signalling (Figure 4C). The time-course was more complex in DMS, with peaks following both the second-step cue and second-step port entry, though the former only survived multiple comparison correction across timepoints in the dLight data (Figure S7). Chosen action value had a strong positive influence on activity in all 3 regions around the time of the choice, which then reversed when the second-step was revealed, again consistent with RPE. These data confirm that value information, which respected the task causal and correlation structure, strongly influenced dopamine responses to both the first-step action and second-step choice.

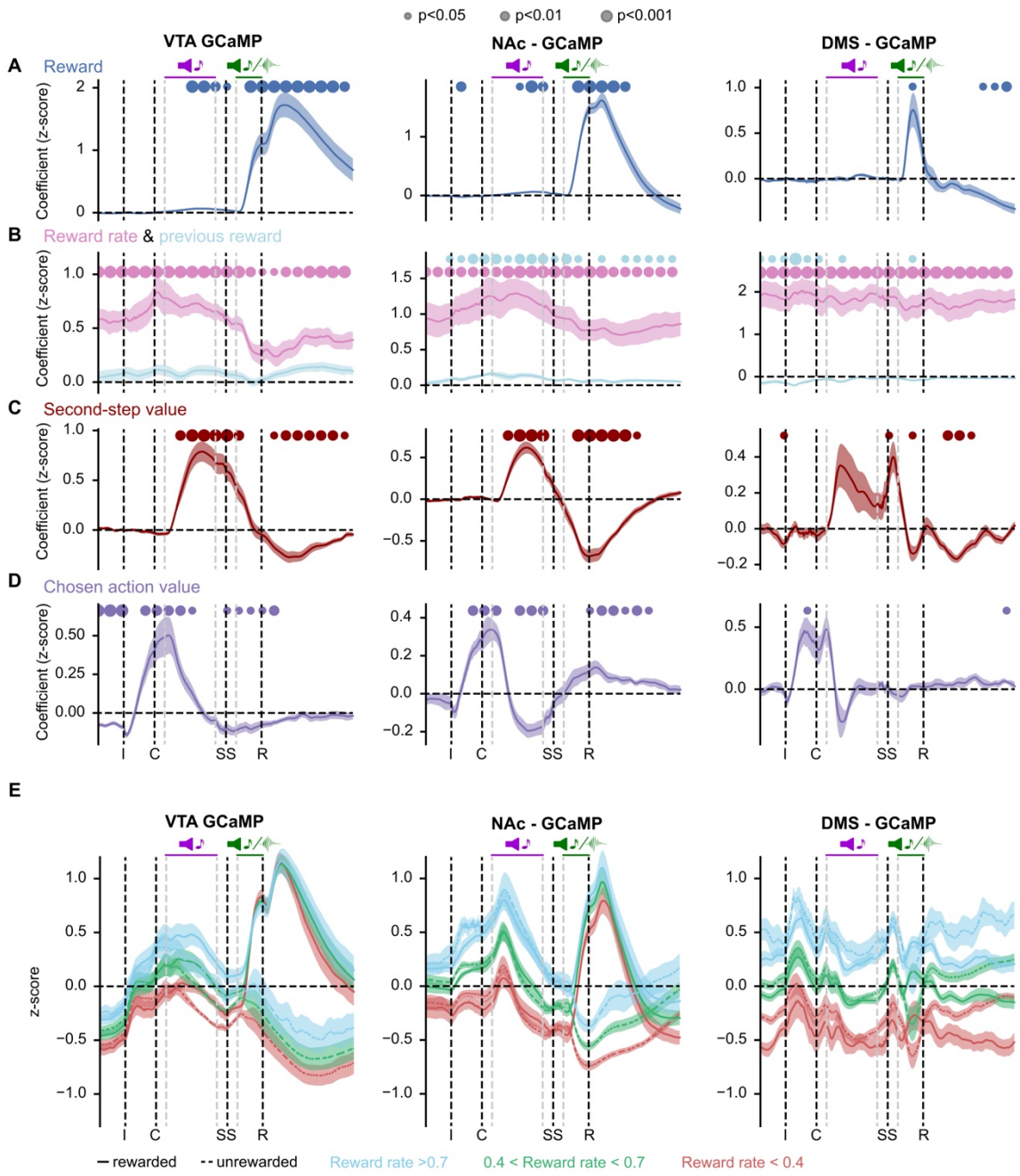


Figure 4. Dissociable influence of value and reward rate on dopamine. **A-D)** Coefficients from linear regression predicting dopamine activity using decision variables from RL model, showing cross-subject mean and standard error (shaded area), dots indicate statistical significance using two-sided t-test comparing the cross-subject distribution against 0 and Benjamini-Hochberg multiple comparison correction, small dot: $p < 0.05$, medium dot: $p < 0.01$, large dot: $p < 0.001$. For predictors: **A)** Current trial outcome. **B)** Recent reward rate (exponential moving average with tau = 10 trials) (pink) and previous trial outcome (blue). **C)** Behavioural model value of second-step reached on current trial. **D)** Behavioural model value of chosen first-step action. **E)** Mean-z-score activity split by rewarded/unrewarded trials and recent reward rate. Blue – high reward rate > 0.7 rewards/trial (exponential moving average with tau = 10 trials). Green, medium reward rate (between 0.4 and 0.7). Red, low reward rate (less than 0.4).

241

242 As both reward rate and value have been shown to correlate with dopamine activity[5,38], we
243 additionally included recent reward rate as a predictor, calculated as an exponential moving
244 average with time constant of 10 trials (qualitatively similar results were obtained using time
245 constants of 5-15 trials, data not shown). Reward rate had a strong positive influence on
246 dopamine in all 3 regions (Figure 4B,E), accounting for separate variance from the prominent
247 influence of action and state values. Unlike the influence of value and reward, which were
248 tightly time-locked to when information became available, reward rate positively influenced
249 activity at all timepoints, with only weak modulation across the trial. This separate influence
250 of value and reward rate was also present in striatal dopamine concentrations (Figure S7).

251 These data show that while action and state values influenced phasic dopamine activity in a
252 manner consistent with RPE signalling, recent reward rate explained separate variance in
253 dopamine activity on a slow timescale.

254

255 *Dopamine stimulation does not recapitulate natural rewards despite reinforcing preceding actions*

256 The above data demonstrate that dopamine reported RPE and reward information, but
257 cannot establish whether this signal was causally responsible for updating mice's choice
258 policy. To assess this, we therefore manipulated dopamine activity in a new cohort of mice
259 expressing either channelrhodopsin (ChR2, N=7) or a control fluorophore (EYFP, N=5) (Figure
260 5A) in VTA dopamine neurons. We verified that our stimulation parameters (5 pulses, 25 Hz,
261 ~8-10 mW power) were sufficient to promote and maintain intracranial self-stimulation in the
262 ChR2 but not YFP groups ($p = 0.011$, two-sided t-test) using an assay where optical stimulation
263 was delivered contingent on nose-poking in a different context from the two-step task (Figure
264 5B).

265 We then examined the effect on two-step task behaviour of optogenetic stimulation at two
266 different timepoints – (i) after the first-step choice, at the time of second-step cue onset, or
267 (ii) at outcome cue presentation (Figure 5C) (25% stimulated trials, stimulation timepoint
268 fixed for each session, counterbalanced across sessions and mice, N=6-8 sessions per mouse
269 and stimulation type).

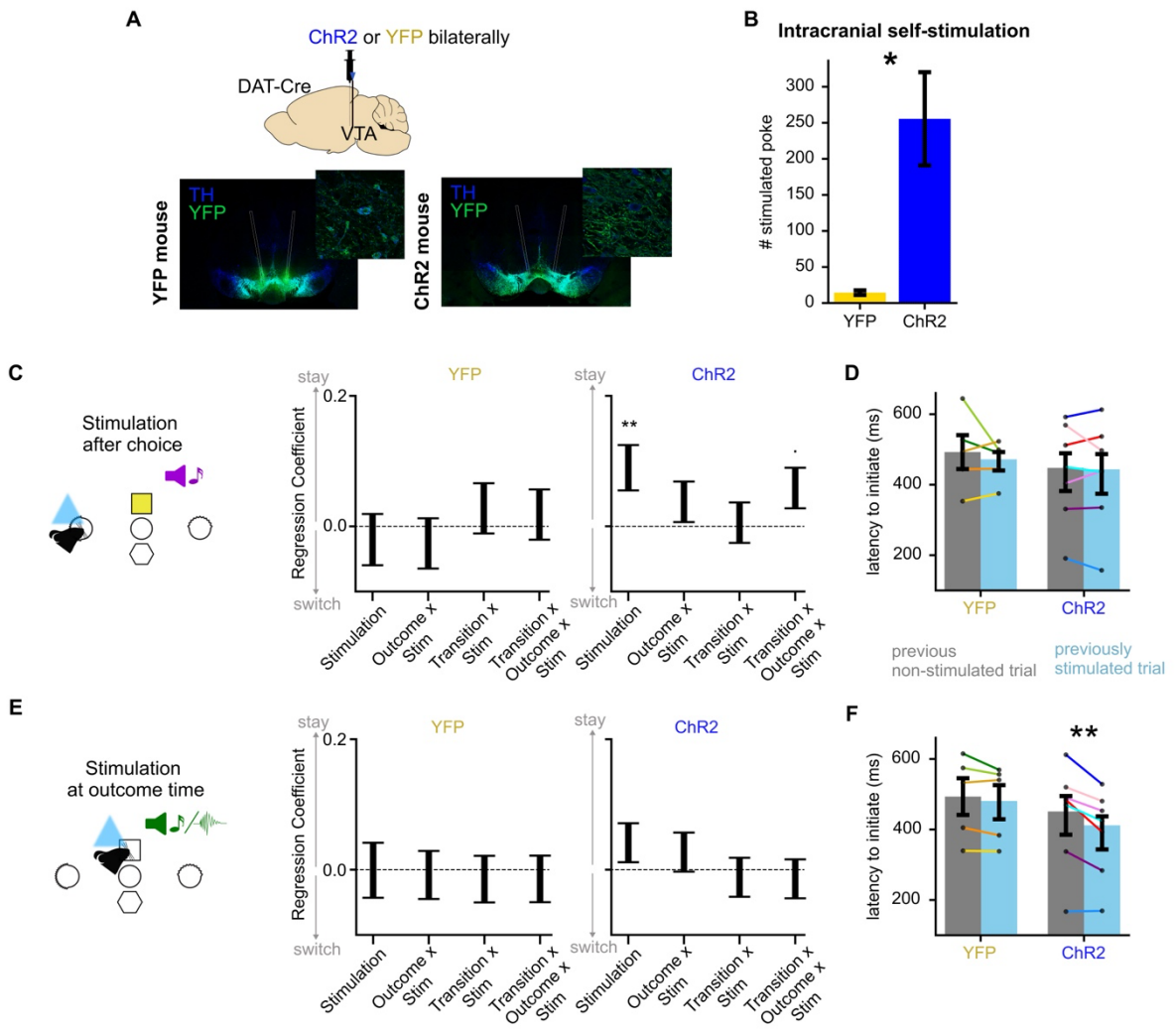


Figure 5. Dopamine stimulation does not recapitulate natural reward effects A) Top: Injection and fibre placement schematics. Bottom: Histology from a YFP and a ChR2 mouse, stained for TH (Tyrosine Hydroxylase) and YFP. **B)** Intracranial self-stimulation. Mean number of entries to a nose-poke port delivering stimulation in the YFP vs the ChR2 group. Error bars show cross subject standard error. *p<0.05 two-sided t-test. **C-D)** Effect of optogenetic stimulation 200ms after choice in the control (YFP) and ChR2 group. **E-F)** Effect of optogenetic stimulation at time of the trial outcome cue. **C,E)** Mixed effects logistic regression predicting repeating the same choice on the next trial using the regression model from Figure 2C, with additional regressors modelling the effect of stimulation and its interaction with trial events. Only regressors showing stimulation effects are shown here, see Figure S8 for the full model. •p<0.1, **p<0.005 **D-F)** Mean latency to initiate trial following stimulated and non-stimulated trials. Dots indicate individual animals, error bars show standard error. **p<0.01, two-sided paired t-test. N: YFP: 5 animals, 13,079 trials (choice-time stimulation sessions), 15,419 trials (outcome-time stimulation sessions); ChR2: 7 animals, 20,817 trials (choice-time stimulation sessions), 23,051 trials (outcome-time stimulation sessions).

In the ChR2 group, stimulating dopamine neurons after the first-step choice was reinforcing; it significantly increased the probability of repeating that choice on the next trial (p=0.0095, mixed effect logistic regression, Figure 5C). This stimulation did not significantly interact with

either the transition or outcome in its effect on next trial choice (all p>0.059). Strikingly, stimulating dopamine neurons at the time of trial outcome had no significant influence on the subsequent choice (Figure 5E); it did not reinforce the preceding first-step choice (effect of stimulation: p=0.17), nor did it act like a reward by interacting with the state transition (p=0.71), nor with the trial outcome to influence on the subsequent choice (p=0.37), nor the two interacting (i.e. outcome interacting with transition to influence subsequent choice: p=0.64). No effect of stimulation was found in the YFP group for either of the stimulation types (all p>0.47).

By contrast, while stimulation after the first step choice had no effect on the latency to initiate the next trial (p=0.74), stimulation at outcome time significantly reduced the latency to initiate the next trial (Figure 5F, p=0.0055, two-sided paired t-test). Again, there was no effect on this latency in the YFP group (p=0.46, stimulation after choice; p=0.16, stimulation at outcome time).

Therefore, while stimulation of VTA dopamine neurons was sufficient to reinforce immediately preceding actions, it completely failed to recapitulate the behavioural consequences of natural rewards at outcome, despite rewards driving the largest dopamine signals observed in the task.

Discussion

We recorded and manipulated dopamine activity in a sequential decision task where rewards do not simply reinforce preceding actions, but rather affect subsequent choices in a manner that respects the task structure. This revealed a striking asymmetry between reporting of RPEs and updating of policy by dopamine. Moment-by-moment dopamine fluctuations were strongly influenced by both the value of actions taken and states encountered, and reward delivery or omission, consistent with RPE signalling. However, while rewards strongly influenced subsequent choices, stimulating dopamine neurons at the time of trial outcome had no effect on animal's next choice. This was despite positive controls verifying that, in the same subjects, the same stimulation was sufficient to (a) reinforce a choice when delivered shortly after it, (b) induce robust self-stimulation in a different context, and (c) speed subsequent trial re-initiation when stimulating at outcome.

Dopamine in inference guided choice

These data are surprising given the tight interaction posited by RL theory between RPE and policy: unpredicted rewards evoke RPEs which update the values of states and actions, and hence shape future policy. However, they can be reconciled with RL theory by recognizing that the task is only partially observable: there are aspects of the task state that cannot be directly observed but which nonetheless can influence observable events (e.g., rewards). Rewards on individual trials (observable events) were determined by underlying reward probabilities, which underwent occasional reversals. The configuration of the reward probabilities on any trial therefore constitutes a latent or unobservable feature of the task's state. Importantly, though not directly observable, it can nonetheless be *inferred* by observing where rewards are obtained[11,12]. Crucially, that means that rewards therefore not only carry information about the affective value, but also about the current *state* of the task. A subject who can infer the latent state of the reward probabilities can learn separate values for actions and trial events conditioned on it. State information provided by task rewards can then influence future choices by changing subjects' beliefs about the latent state, and hence which values are in play[11,14]. In this account, subjects adapt to reversals by changing their beliefs about the current state of the task, not by changing the values associated with each state.

Several lines of evidence support the notion that the mice were solving this task using state inference. First, the best fitting behavioural model used a single variable updated in opposite directions by rewards at the up and down port, rather than learning values for the two second-step states independently. This can be thought of as representing subject's belief about whether the reward probabilities are currently high at the up or down port. It is consistent with state inference because it incorporates knowledge that rewards at the two locations are not independent but rather controlled by a single latent variable. Unexpectedly, only rewards but not reward omissions appear to update this variable, a conclusion supported both by model comparison and the lagged regression analysis (Figure 1F). This is not optimal as both rewards and reward omissions were informative, and might reflect either cognitive limitations or divergence between task statistics and those of ethological foraging.

Second, our photometry data showed that rewards obtained in one second-step state (e.g. up-active) *negatively* affected the dopamine response upon reaching the *other* second step state (down-active) on the subsequent trial (Figure 3B). This is inconsistent with learning independent values for the two states, as reward in one state would then not affect the value of the other, but is consistent with the inference that reward obtained at one second-step state makes obtaining reward at the other less likely.

Third, our manipulation results are consistent under the assumption that dopamine neurons carry RPEs evoked by rewards, used to update the values of states and actions, but not the state prediction errors used to update beliefs[39]. Stimulating dopamine at the time of trial outcome will increase the value of that second-step, but this change in second-step value would only influence the next trial choice if subjects were doing prospective model-based evaluation at the first-step. However, if subjects learn cached values conditioned on the inferred latent state, there is no advantage to doing costly model-based planning, because the cached values will generate very similar behaviour[11]. Specifically, cached values learnt conditioned on the latent state will be high for the first-step action that commonly leads to the second-step port with high reward probability. By contrast, stimulating dopamine immediately after choice will reinforce that choice by directly increasing the cached value of the chosen action.

These data build on previous evidence for inference underlying behavioural flexibility[12–16,18], and informing dopamine responses[15,22,24,27,40], but extend them by both recording and manipulating dopamine during inference-guided adaptive decisions, revealing a disconnect between value reporting and policy updating. An inference-based perspective helps explain previous observations that are surprising under a pure value learning account of flexible behaviour. Two studies using probabilistic reward-guided tasks, found that optogenetic inhibition of dopamine neurons on single trials biased animals against the chosen option on the subsequent trial[5,6], interpreted as evidence that task behaviour was dopamine-dependent and dopamine inhibition acted as a negative RPE. However, in both cases, the reinforcing effect of rewards was unaffected by inhibiting dopamine neurons. While this is inconsistent with dopaminergic RPEs mediating reward effects, it is consistent with the inference-based account above. Similarly, pharmacological manipulation of dopamine signalling in reward

363 guided decision tasks often has limited or no effect on learning[12,41–43], suggesting a dopamine-
364 independent influence of rewards on policy.

365 This does not imply dopamine is not involved in such tasks. Pharmacological dopamine
366 manipulations show effects on choice consistency and task engagement even where they
367 show limited or no effect on trial-by-trial learning[42–44]. Dopaminergic RPEs likely also drive
368 value learning during initial task acquisition. As representations of observable events are
369 available immediately, but those of latent states will emerge only gradually with task
370 experience, dynamic arbitration must occur between values learnt over observable and
371 inferred features. Normatively, this should be determined by their predictive accuracy[45],
372 perhaps implemented via synaptic learning rules which track uncertainty[46]. As values
373 conditioned on correctly inferred latent states are stable across changes in those states, they
374 can presumably be learnt with higher accuracy than values over observable features, which
375 change when latent states change. This suggests that over task acquisition, control will
376 transfer from values learnt over observable features to those over latent features, and hence
377 the mechanism of behavioural flexibility from value-learning to inference.

378 Recording, manipulation and neuroimaging studies implicate orbitofrontal, medial prefrontal
379 and medial temporal lobe regions in learning and representing latent structure[13,14,53,15,16,47–
380 52]. This may reflect their position near the top of hierarchically organised predictive models,
381 which learn to minimise sensory prediction errors by inferring their latent causes[54–56].
382 Another possible learning mechanism is meta RL, where plasticity in a recurrent network –
383 putatively frontal cortex – on a slow timescale, driven by reward prediction errors,
384 implements network dynamics that adapt to environment changes on a fast timescale[57].

385

386 *RPE, value, and reward rate*

387 Recent work has questioned the relative influence of value and RPE on dopamine activity and
388 striatal concentrations[4,5,38]. A key signature of RPE is that the value of encountered states
389 contributes at two timepoints with opposite signs – first positively, when the state is entered,
390 then negatively, on transition to the next state[35–37]. We observed this biphasic influences of
391 the values of both first-step actions and second-step states in both calcium activity in VTA,
392 NAc and DMS (Figure 4C,D), and in dopamine concentrations in NAc and DMS (Figure S7). This

biphasic pattern was most prominent in NAc, with VTA and DMS signals both showing substantially larger initial positive modulation than later negative modulation. Intriguingly, when evaluating only the influence of the most recent outcome on second-step state value, rather than the extended history, the biphasic pattern was prominent in NAc but absent in VTA (Figure 3B). Differences between the VTA and striatal signals could reflect local modulation of terminals, for example by cholinergic neurons[58], or alternatively the VTA signal may be influenced by calcium activity in dendrites that is partially decoupled from spiking due to somatic inhibition.

In addition, we observed the average reward rate over recent trials had a strong positive influence on dopamine activity in all regions (Figure 4B, S6). Importantly, this accounted for an entirely separate portion of variance from the value of trial events. Value-related signals were tightly time-locked to specific events, whereas reward rate had sustained effects across the trial from before initiation to after outcome. This appears broadly consistent with theoretical proposals that tonic dopamine represents average reward rate, which acts as the opportunity cost of time in average-reward RL[59]. Correlation between reward rate and NAc dopamine concentrations is consistent with recent reports[5,38], but that with VTA calcium is unexpected given previous reports of no correlation with VTA spikes[38].

Conclusions

Our findings emphasise that flexible behaviour involves two processes operating in parallel: inference about the current state of the world, and evaluation of those states and actions taken in them. The involvement of dopamine in updating values has rightly been a major focus of accounts of flexible decision making. However, in the structured environments common both in the lab and real world, it is only half the picture, as subjects can adapt to change by updating their beliefs about the world's state. We showed here that although such belief updates strongly influence dopamine activity, this relationship is not reciprocal, as stimulating dopamine failed to reproduce natural reward effects. A pressing direction for future work is to understand the interplay between these two processes in flexibly adapting to a changing world.

Acknowledgements

This research was funded by the Wellcome (202831/Z/16/Z: MEW; 214314/Z/18/Z: MEW, TA; WT096193AIA: TA; 215198/Z/19/Z: MBP). 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. We thank Marios Panayi for assistance with the mixed effects modelling, Tim Behrens for discussions about the work, and Peter Dayan and Armin Lak for comments on the manuscript.

Author Contributions

TA and MEW conceived the project and, with MBP, designed the experiments. MBP and TA collected the data. MBP and TA analyzed the data. All authors wrote the manuscript.

Conflict of Interest

The authors have no competing interests to declare.

437 **References**

438 1. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic
439 dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–47
440 (1996).

441 2. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward.
442 *Science* **275**, 1593–9 (1997).

443 3. Eshel, N. *et al.* Arithmetic and local circuitry underlying dopamine prediction errors.
444 *Nature* **525**, 243–6 (2015).

445 4. Kim, H. R. *et al.* A Unified Framework for Dopamine Signals across Timescales. *Cell*
446 **183**, 1600-1616.e25 (2020).

447 5. Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. *Nat. Neurosci.* **19**,
448 117–26 (2016).

449 6. Parker, N. F. *et al.* Reward and choice encoding in terminals of midbrain dopamine
450 neurons depends on striatal target. *Nat. Neurosci.* **19**, 845–54 (2016).

451 7. Kravitz, A. V, Tye, L. D. & Kreitzer, A. C. Distinct roles for direct and indirect pathway
452 striatal neurons in reinforcement. *Nat. Neurosci.* **15**, 816–8 (2012).

453 8. Kim, K. M. *et al.* Optogenetic mimicry of the transient activation of dopamine neurons
454 by natural reward is sufficient for operant reinforcement. *PLoS One* **7**, e33612 (2012).

455 9. Steinberg, E. E. *et al.* A causal link between prediction errors, dopamine neurons and
456 learning. *Nat. Neurosci.* **16**, 966–73 (2013).

457 10. Bartolo, R. & Averbeck, B. B. Inference as a fundamental process in behavior. *Curr.*
458 *Opin. Behav. Sci.* **38**, 8–13 (2021).

459 11. Akam, T., Costa, R. & Dayan, P. Simple Plans or Sophisticated Habits? State, Transition
460 and Learning Interactions in the Two-Step Task. *PLoS Comput. Biol.* **11**, e1004648
461 (2015).

462 12. Costa, V. D., Tran, V. L., Turchi, J. & Averbeck, B. B. Reversal learning and dopamine: a
463 bayesian perspective. *J. Neurosci.* **35**, 2407–16 (2015).

464 13. Vertechi, P. *et al.* Inference-Based Decisions in a Hidden State Foraging Task:

Differential Contributions of Prefrontal Cortical Areas. *Neuron* **106**, 166-176.e6 (2020).

14. Hampton, A. N., Bossaerts, P. & O'Doherty, J. P. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.* **26**, 8360–7 (2006).

15. Wimmer, G. E., Daw, N. D. & Shohamy, D. Generalization of value in reinforcement learning by humans. *Eur. J. Neurosci.* **35**, 1092–104 (2012).

16. Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M. & Behrens, T. E. J. Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron* **109**, 713-723.e7 (2021).

17. Izquierdo, A., Suda, R. K. & Murray, E. A. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *J. Neurosci.* **24**, 7540–8 (2004).

18. Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. & Akam, T. Complementary Task Representations in Hippocampus and Prefrontal Cortex for Generalising the Structure of Problems. *bioRxiv* 2021.03.05.433967 (2021). doi:10.1101/2021.03.05.433967

19. Mackintosh, N. J., McGonigle, B., Holgate, V. & Vanderver, V. Factors underlying improvement in serial reversal learning. *Can. J. Psychol.* **22**, 85–95 (1968).

20. Bond, A. B., Kamil, A. C. & Balda, R. P. Serial reversal learning and the evolution of behavioral flexibility in three species of North American corvids (Gymnorhinus cyanocephalus, Nucifraga columbiana, Aphelocoma californica). *J. Comp. Psychol.* **121**, 372–9 (2007).

21. Gallistel, C. R., Mark, T. A., King, A. P. & Latham, P. E. The rat approximates an ideal detector of changes in rates of reward: implications for the law of effect. *J. Exp. Psychol. Anim. Behav. Process.* **27**, 354–72 (2001).

22. Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.* **104**, 1068–76 (2010).

494 23. Nakahara, H., Itoh, H., Kawagoe, R., Takikawa, Y. & Hikosaka, O. Dopamine neurons
495 can represent context-dependent prediction error. *Neuron* **41**, 269–80 (2004).

496 24. Starkweather, C. K., Babayan, B. M., Uchida, N. & Gershman, S. J. Dopamine reward
497 prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–
498 589 (2017).

499 25. Engelhard, B. *et al.* Specialized coding of sensory, motor and cognitive variables in
500 VTA dopamine neurons. *Nature* **570**, 509–513 (2019).

501 26. Lak, A. *et al.* Neural basis of learning guided by sensory confidence and reward value.
502 *bioRxiv* 411413 (2019). doi:10.1101/411413

503 27. Gershman, S. J. & Uchida, N. Believing in dopamine. *Nat. Rev. Neurosci.* **20**, 703–714
504 (2019).

505 28. Tsutsui-Kimura, I. *et al.* Distinct temporal difference error signals in dopamine axons
506 in three regions of the striatum in a decision-making task. *Elife* **9**, (2020).

507 29. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based
508 influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–15
509 (2011).

510 30. Deserno, L. *et al.* Ventral striatal dopamine reflects behavioral and neural signatures
511 of model-based control during sequential decision making. *Proc. Natl. Acad. Sci. U. S.*
512 *A.* **112**, 1595–600 (2015).

513 31. Colas, J. T., Pauli, W. M., Larsen, T., Tyszka, J. M. & O'Doherty, J. P. Distinct prediction
514 errors in mesostriatal circuits of the human brain mediate learning about the values
515 of both states and actions: evidence from high-resolution fMRI. *PLoS Comput. Biol.*
516 **13**, e1005810 (2017).

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

520 33. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during
521 locomotion and reward. *Nature* **535**, 505–10 (2016).

522 34. Akam, T. *et al.* The Anterior Cingulate Cortex Predicts Future States to Mediate

523 Model-Based Action Selection. *Neuron* **109**, 149-163.e7 (2021).

524 35. Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. Associative
525 learning of social value. *Nature* **456**, 245–9 (2008).

526 36. Rutledge, R. B., Dean, M., Caplin, A. & Glimcher, P. W. Testing the reward prediction
527 error hypothesis with an axiomatic model. *J. Neurosci.* **30**, 13525–36 (2010).

528 37. Niv, Y., Edlund, J. A., Dayan, P. & O'Doherty, J. P. Neural prediction errors reveal a
529 risk-sensitive reinforcement-learning process in the human brain. *J. Neurosci.* **32**,
530 551–62 (2012).

531 38. Mohebi, A. *et al.* Dissociable dopamine dynamics for learning and motivation. *Nature*
532 **570**, 65–70 (2019).

533 39. Akam, T. & Walton, M. E. What is dopamine doing in model-based reinforcement
534 learning? *Curr. Opin. Behav. Sci.* **38**, 74–82 (2021).

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

537 41. Grogan, J. P. *et al.* Effects of dopamine on reinforcement learning and consolidation
538 in Parkinson's disease. *Elife* **6**, (2017).

539 42. Verharen, J. P. H., Adan, R. A. H. & Vanderschuren, L. J. M. J. Differential contributions
540 of striatal dopamine D1 and D2 receptors to component processes of value-based
541 decision making. *Neuropsychopharmacology* **44**, 2195–2204 (2019).

542 43. Korn, C. *et al.* Distinct roles for dopamine clearance mechanisms in regulating
543 behavioral flexibility. *bioRxiv* 823401 (2021). doi:10.1101/823401

544 44. Lee, E., Seo, M., Dal Monte, O. & Averbeck, B. B. Injection of a dopamine type 2
545 receptor antagonist into the dorsal striatum disrupts choices driven by previous
546 outcomes, but not perceptual inference. *J. Neurosci.* **35**, 6298–306 (2015).

547 45. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal
548 and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–11
549 (2005).

550 46. Aitchison, L. *et al.* Synaptic plasticity as Bayesian inference. *Nat. Neurosci.* **24**, 565–

551 571 (2021).

552 47. Schuck, N. W., Cai, M. B., Wilson, R. C. & Niv, Y. Human Orbitofrontal Cortex
553 Represents a Cognitive Map of State Space. *Neuron* **91**, 1402–1412 (2016).

554 48. Klein-Flügge, M. C., Wittmann, M. K., Shpektor, A., Jensen, D. E. A. & Rushworth, M. F.
555 S. Multiple associative structures created by reinforcement and incidental statistical
556 learning mechanisms. *Nat. Commun.* **10**, 4835 (2019).

557 49. Bradfield, L. A., Dezfouli, A., van Holstein, M., Chieng, B. & Balleine, B. W. Medial
558 Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task
559 Situations. *Neuron* **88**, 1268–1280 (2015).

560 50. Starkweather, C. K., Gershman, S. J. & Uchida, N. The Medial Prefrontal Cortex Shapes
561 Dopamine Reward Prediction Errors under State Uncertainty. *Neuron* **98**, 616-629.e6
562 (2018).

563 51. Jocham, G. *et al.* Reward-Guided Learning with and without Causal Attribution.
564 *Neuron* **90**, 177–90 (2016).

565 52. Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H. & Rushworth, M. F. S.
566 Separable learning systems in the macaque brain and the role of orbitofrontal cortex
567 in contingent learning. *Neuron* **65**, 927–39 (2010).

568 53. Bartolo, R. & Averbeck, B. B. Prefrontal Cortex Predicts State Switches during Reversal
569 Learning. *Neuron* **106**, 1044-1054.e4 (2020).

570 54. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate
571 cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).

572 55. Friston, K. Hierarchical models in the brain. *PLoS Comput. Biol.* **4**, e1000211 (2008).

573 56. Harris, J. A. *et al.* Hierarchical organization of cortical and thalamic connectivity.
574 *Nature* **575**, 195–202 (2019).

575 57. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat.*
576 *Neurosci.* **21**, 860–868 (2018).

577 58. Threlfell, S. & Cragg, S. J. Dopamine signaling in dorsal versus ventral striatum: the
578 dynamic role of cholinergic interneurons. *Front. Syst. Neurosci.* **5**, 11 (2011).

59. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology (Berl).* **191**, 507–20 (2007).

60. Akam, T. *et al.* pyControl: Open source, Python based, hardware and software for controlling behavioural neuroscience experiments. *bioRxiv* 2021.02.22.432227 (2021). doi:10.1101/2021.02.22.432227

61. Lopes, G. *et al.* Bonsai: an event-based framework for processing and controlling data streams. *Front. Neuroinform.* **9**, 7 (2015).

62. Akam, T. & Walton, M. E. pyPhotometry: Open source Python based hardware and software for fiber photometry data acquisition. *Sci. Rep.* **9**, 3521 (2019).

63. Singmann, H., Bolker, B., Westfall, J. & Aust, F. afex': analysis of factorial experiments. R package. (2018).

METHODS

Subjects

All experiments were performed in line with the UK Animal (Scientific Procedure) Act 1986 and in accordance with the University of Oxford animal use guidelines. 12 dopamine transporter (DAT)-Cre heterozygous mice (7 females and 5 males) were used for the GCaMP photometry recordings, 6 WT C57BL/6 mice (3 females and 3 males) for the dLight recordings, and 12 DAT-Cre mice (YFP: 2 females and 3 males; ChR2: 4 females and 3 males) for the optogenetic manipulation experiment. Mice were aged 8-16 weeks at the start of behavioural training. Animals were typically housed in groups of 2-4 throughout training and testing.

Data and code availability

All data and code used in the manuscript will be made available in a public repository from acceptance of the manuscript.

Behavioural setup

The task was run in custom-built 12x12 cm operant boxes (design files at https://github.com/pyControl/hardware/tree/master/Behaviour_box_small) controlled using pyControl[60]. Five nose poke ports were located on the back wall of the boxes; a central initiation port flanked by two choice ports 4 cm to the left and right and two second-step ports located 1.6 cm above and below the central poke. The second-step ports each had a solenoid to deliver water rewards. A speaker located above the ports was used to deliver auditory stimuli. Video was acquired from an FLIR Flea 3 camera positioned above each setup using a Bonsai based workflow (https://github.com/ThomasAkam/Point_Grey_Bonsai_multi_camera_acquisition) [61].

Behavioural task and training

The behavioural task was adapted from the human two-step task[29]. Each trial started with the central initiation port lighting up. Subjects initiated the trial by poking the illuminated port, which caused the choice ports to illuminate. On free-choice trials (75% of trials), both

621 the left and right port lit up, allowing subjects to choose, while on forced-choice trials only
622 one randomly selected choice port lit up, forcing animals to select that specific port. Poking
623 a choice port was followed, after a 200ms delay, by the second-step port lighting up and 1s
624 presentation of one of two auditory cues (either a 5- or 12-kHz tone depending on whether
625 the top or bottom second-step became active). Each choice port was commonly (80% trials)
626 associated with transitioning to one second-step state (up or down) and rarely (20% trials) to
627 the other. The second-step port only became responsive to pokes after cue offset. Poking the
628 second-step port triggered a 200ms delay, after which a 500ms auditory cue signalled
629 whether the trial was rewarded or not (same 5- or 12-kHz tone pulsed with at 10Hz on
630 rewarded trials, white noise on unrewarded trials). Reward was delivered at the offset of this
631 cue. To ensure mice knew when they had made a nose poke, a click sound was presented
632 whenever the subject poked a port that was active (e.g., the initiation port during the
633 initiation state).

634 Reward probabilities for the two second-step ports changed in blocks. In non-neutral blocks,
635 one second-step port had 80% reward probability and the other 20%, while in neutral blocks
636 both second-step ports were rewarded with 50% probability. Block transitions from non-
637 neutral blocks were triggered 5 to 15 trials after mice crossed a threshold of 75% "correct"
638 choices (i.e., choosing the higher reward probability), computed as the exponential moving
639 average with time constant of 8 free choice trials. Transitions from neutral blocks were
640 triggered after a 20-30 trials. An inter-trial interval of 2-4 s duration started once the subject
641 remained out of the second-step port for 250ms after the trial outcome.

642

643 *Training*

644 Behavioural training took 4-6 weeks. Animals were put on water restriction 48 hours before
645 starting training, and received 1 hour ad lib water access in their home cage 24 hours before
646 starting training. On training days (typically 6 days/week) animals usually received all their
647 water from the task, but were topped up outside the task as necessary to maintain a body
648 weight >85% of their pre restriction baseline. On days off from training mice received 1 hour
649 ad lib water access in their home cage. Water reward size was decreased from 15ul to 4ul
650 across training in order to increase the number of trials performed and blocks experienced
651 on each session.

Training consisted of multiple stages with increasing complexity (Table 1). At the first training stage 1.1, only the second-step ports were exposed, with all other ports covered. Second-step ports were illuminated in a pseudorandom order with inter-trial interval of 2-4 seconds. Poking an illuminated port delivered reward with 100% probability, with no auditory cues. When animals obtained >50 rewards on this stage, transitioned to stage 1.2, where the auditory cues for second-step state and reward were introduced. Once animals obtaining >70 rewards in a session, they were switched to stage 2 on the next session. At stage 2, the choice ports were introduced, but all trials were forced choice, such that only one choice port lit up on each trial. Mice were switched to stage 3 when they obtained >70 rewards on a single session. At stage 3, the initiation poke was introduced, and when they obtained >70 rewards on a single session, they transitioned to stage 4 on the next training session. Stage 4 consisted of multiple substages where progressively more free choice trials were introduced, and the reward probabilities gradually changed until reaching the final task parameters. Mice were transitioned to next substage after two training sessions of 45min or a single 90min session, until they reached substage 4.6. Subjects were only transition to the final stage (full task) when they completed at least 5 blocks in a single session.

STAGE	START POKE	CHOICE POKES	REWARD POKES	TASK STRUCTURE
1.1			X	Reward probability 100%
1.2				Introduce auditory cues Reward probability 100%
2		X	X	Transition probability 80/20% Reward probability 100%
3	X	X	X	Transition probability 80/20% Reward probability 100%
	x	x	x	Start introducing free choice trials and reward probabilities
SUBSTAGE	**Free choice trials**	**Non-neutral blocks reward probabilities** (high / low)	**Neutral block reward probability**	**Block change** (independent / dependent on behaviour)
4.1	0%	0.9/0.7	0.8	Independent – 20-30 trials blocks
4.2	25%	0.9/0.5	0.7	Independent – 20-30 trials blocks
4.3	25%	0.9/0.3	0.6	Independent – 20-30 trials blocks
4.4	25%	0.9/0.1	0.5	Independent – 20-30 trials blocks
4.5	50%	0.9/0.1	0.5	Dependent
4.6	75%	0.9/0.1	0.5	Dependent
4.7 – FINAL	75%	0.8/0.2	0.5	Dependent

Table 1. Description of training stages and task contingencies

673 **Surgery**

674 Mice were anaesthetised with isoflurane (3% induction, 0.5-1% maintenance), and injected with
675 buprenorphine (0.08 mg/kg), meloxicam (5 mg/kg) and glucose saline (0.5ml). Marcaine (max 2
676 mg/kg) was injected into the scalp and before placing mice into the stereotaxic frame. Mice were
677 maintained at ~37 degrees with rectal probe and heating blanket (Harvard Apparatus). Surgery
678 proceeded as described below for the different experiments. After surgery, mice were given
679 additional doses of meloxicam each day for 3 days after surgery, and were monitored carefully for 7
680 days post-surgery.

681

682 *GCaMP photometry*

683 Mice were intracranially injected with 1ul of saline containing
684 AAV1.Syn.Flex.GCaMP6f.WPRE.SV40 (titre: 6.22 x 10^{12} vg/ml, Penn Vector Core) and
685 AAV1.CAG.Flex.tdTomato.WPRE.bGH (AllenInstitute864) (titre: 1.535 x 10^{12} vg/ml, Penn
686 Vector Core) at 2nl/s in VTA (antero-posterior (AP): -3.3, medio-lateral (ML): ±0.4, dorso-
687 ventral (DV): -4.3 from bregma) in one hemisphere for mesolimbic dopamine recordings in
688 VTA and NAc and VTA/SNc (AP: -3.1, ML: ±0.9, DV: -4.2 from bregma) in the other hemisphere
689 for DMS recordings. Three 200 µm diameter ceramic optic fibres were implanted chronically
690 in each animal in VTA (AP: -3.3, ML: ±0.4, DV: -4.3 from bregma) and NAc (AP: +1.4, ML: ±0.8,
691 DV: -4.1 from bregma) in the same hemisphere, and DMS (AP: +0.5, ML: ±1.5-1.7, DV: -2.6
692 from bregma) in the contralateral hemisphere.

693

694 *dLight photometry*

695 Mice were intracranially injected with 500 nl of saline containing pAAV5-CAG-dLight1.1 (titre:
696 1.4 x 10^{12} vg/ml, addgene) and pssAAV-2/5-hSyn1-chI-tdTomato-WPRE-SV40p(A) (titre: 4.9 x
697 10^{11} vg/ml, ETH Zurich) at 2nl/s in NAc (AP: +1.4, ML: ±0.8, DV: -4.1 from bregma) and DMS
698 (AP: +0.5, ML: ±1.5/1.7, DV: -2.6 from bregma) in opposite hemispheres. Two 200 µm
699 diameter ceramic optic fibres were implanted chronically in the injection sites.

700

701 *Optogenetic stimulation*

Mice were injected bilaterally with 500 nl/hemisphere of saline containing either AAV2-EF1a-DIO-EYFP (titre: >1 x 10^{12} vg/ml, UNC Vector Core) (YFP group) or rAAV2/Ef1a-DIO-hchR2(E123t/T159C)-EYFP (titre: 5.2 x 10^{12} vg/ml, UNC Vector Core) (ChR2 group) at 2 nl/s in VTA (AP: -3.3, ML: ±0.4, DV: -4.3 from bregma). Two 200 µm diameter ceramic optic fibres were implanted chronically in the injection sites at a 10-degree angle.

Histology

Mice were terminally anesthetised with sodium pentobarbital and transcranial perfused with saline then 4% PFA solution. 50 µm coronal brain slices were cut, covering striatum and VTA, and immunostained with anti-GFP and anti-TH primary antibodies, and 488 and Cy5 secondary antibodies. An Olympus FV3000 microscope was used to image the slices.

Photometry recordings

Dopamine calcium activity (GCaMP) and release (dLight) were recorded at 130Hz sampling rate using pyPhotometry[62]. The optical system comprised a 465 nm and a 560 nm LED, a 5 port minicube and fibre optic rotary joint (Doric Lenses), and two Newport 2151 photoreceivers. Time division illumination with background subtraction was used to prevent crosstalk between fluorophores due to the overlap of their emission spectra, and changes in ambient light from affecting the signal. Synchronisation pulses from pyControl onto a digital input of the pyPhotometry board were used to synchronise the photometry signal with behaviour[62].

Photometry signals were pre-processed using custom Python code. A median filter (width 5 samples) was first used to remove any spikes due to electrical noise picked up by the photodetectors. To motion-correct the signals, we band-passed the detrended signals between 0.001 Hz and 5 Hz, and used linear regression to predict the GCaMP or dLight signal using the control fluorophore (tdTomato) signal. The predicted signal due to motion was subtracted from the detrended signal. To correct for bleaching of the fibre and fluorophores, signals were detrended using a double exponential fit. Afterwards, a 5Hz low-pass filter was used to denoise the signal. Finally, the pre-processed dopamine signal was z-scored for each session to allow comparison across sessions and animals with different signal intensities.

732 For the GCaMP data, we recorded from 12, 11 and 12 mice in VTA, NAc and DMS, respectively
733 (in one animal, the fibre targeting NAc did not exhibit any GCaMP modulation, which later
734 histological analysis confirmed was caused by the fibre being in the anterior commissure). In
735 DMS, 2 animals – the two with most medial coordinates – were excluded from the main
736 analyses of the effects of reward on subsequent dopamine activity as they presented a
737 negative modulation to reward (Figure S3).

738 For the dLight data, we recorded from 5 and 6 mice in NAc and DMS, respectively (1 subject
739 in NAc did not present any dLight modulation, posterior histological analysis confirmed that
740 the fibre was misplaced into the ventricle).

741 Sessions in which there were large artifacts (large step change in recorded signals) introduced
742 through a malfunctioning of the rotary joint or disconnection of the patch cord from the fibre,
743 or where there was a complete loss of signal on one of the channels due to discharged battery
744 during recording, were excluded. A total of 46 sessions (~9% of the total) were removed from
745 the analyses.

746

747 **Optogenetic manipulation**

748 VTA dopamine neurons were stimulated bilaterally using two 465 nm LEDs (Plexon Plexbright)
749 connected to 200 µm 0.66 NA optic fiber patch cords. All stimulation experiments used
750 stimulation parameters of 5 pulses at 25 Hz, 25% duty cycle, ~8-10mW optical power at the
751 fibre tip.

752 First, as a positive control, we performed an intracranial self-stimulation (ICSS) assay in which
753 mice were presented with either 4 or 2 nose poke ports, one of which triggered optical
754 stimulation when poked. A minimum 1-second delay was imposed between stimulations.
755 Mice were tested on ICSS on 40-60min sessions for 4 days.

756 We then tested the effect of optogenetic manipulation during the two-step task. On each
757 stimulation session, we either stimulated (i) 200ms after the first-step choice at the time of
758 second-step cue onset, or (ii) at outcome-cue onset. Stimulation occurred on 25% of the
759 trials, under the constraints that (i) the trial after stimulation was always a free choice trial,
760 and (ii) there were always at least 2 non-stimulated trials after each stimulation. The
761 stimulation sessions were interspersed with baseline no stimulation sessions (data not

shown). The timing of stimulation was fixed within a session, with the session order counterbalanced across animals (e.g., no stimulation session → second-step cue stimulation session → outcome stimulation session → no stimulation session → outcome stimulation session → second-step cue stimulation session, etc).

Analysis

All behavioural and photometry analysis was performed using custom Python code.

Behavioural logistic regression model

The logistic regression analysis shown in Figure 1E predicted repeating choices (or "staying") as a function of the subsequent trial events, considering only free choice trials, implemented as a mixed effects model using the afex package[63] in R programming language. The model formula was:

$$stay = correct + bias + transition * outcome + (random\ effects|\ subject)$$

For the analysis of two-step optogenetic manipulation (Figure 5C,E), we added stimulation and its interaction with trials events as addition predictors, giving the formula:

$$stay = correct + bias + transition * outcome * stim + (random\ effects||\ subject)$$

We used orthogonal sum-to-zero contrasts, and likelihood ratio test to calculate p-values. The random effects structure used subject as a grouping factor and all random slopes and intercepts that could be included without having a singular fit. This was achieved by removing any random effects fitting to close to zero variance.

The predictors were coded as:

- *Correct*: 3 level categorical variable indicating whether the previous choice was correct, incorrect or in a neutral block.
- *Bias*: Binary categorical variable explaining whether previous choice was left or right.
- *Outcome*: Binary categorical variable indicating whether the previous trial was rewarded or not.
- *Transition*: Binary categorical variable indicating whether previous trial had a common or a rare transition.

- *Stimulation*: Binary categorical variable indicating whether the previous trial was stimulated or not. This regressor was only used for the optogenetic analysis.

For the optogenetic manipulation analysis, we performed separate regressions per group (YFP & ChR2) and stimulation time (after choice & outcome). To ensure stimulated and non-stimulated trials had matching histories, we only included trials where stimulation could have potentially been delivered, i.e., excluding the 2 trials following each stimulation where stimulation was never delivered.

The lagged logistic regression analysis (Figure 1F) assessed how subjects' choices were affected by the history of trial events over the last 12 trials. The regression predicted subjects' probability of choosing left, using the following predictors at lags 1, 2, 3-4, 5-8, 9-12 (where lag 3-4 etc. means the sum of the individual trial predictors over the specified range of lags).

- *Common Transition*: Rewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a common transition and reward, 0 otherwise.
- *Rare Transition*: Rewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a rare transition and reward, 0 otherwise.
- *Common Transition*: Unrewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a common transition and no reward, 0 otherwise.
- *Rare Transition*: Unrewarded at lag n: +0.5/-0.5 if the n'th previous trial was a left/right choice followed by a rare transition and no reward, 0 otherwise.

The lagged regression was fitted separately for each subject as a fixed effects model. Cross subjects mean and standard error for each predictor coefficient were plotted, significance of coefficients was assessed using a two-sided t-test comparing the distribution of the individual subjects' coefficients against zero, and performing Bonferroni multiple comparison correction.

Behavioural models

We evaluated goodness-of-fit to subjects' choices for a set of different reinforcement learning (RL) agents created by combining one or more of the following learning strategies:

Model-free

The model-free strategy updated value $Q_{MF}(c)$ of the chosen action and value $V(s)$ of second-step state as:

$$Q_{MF}(c) \leftarrow (1 - \alpha)Q_{MF}(c) + \alpha((1 - \lambda)V(s) + \lambda r)$$

$$V(s) \leftarrow (1 - \alpha)V(s) + \alpha r$$

Where α is the learning rate, λ is the eligibility trace parameter and r is the outcome.

Model-based

The model-based strategy updated value $V(s)$ of the second-step reached, and both first-step action values $Q_{MB}(a)$ as:

$$V(s) \leftarrow (1 - \alpha)V(s) + \alpha r$$

$$Q_{MB}(a) = \sum_{s} P(s|a)V(s)$$

Where α is the learning rate, r is the outcome, $P(s|a)$ is the transition probability of reaching second-step s after taking action a (i.e. 0.8 or 0.2).

Inference

The inference strategy used a single variable B to track the anticorrelated reward probabilities at the two second-step states, representing subjects' belief about which second-step state currently had high reward probability. This was updated as:

$$B \leftarrow (1 - \alpha_r)B + \alpha_r S \qquad \text{if previous trial rewarded}$$

$$B \leftarrow (1 - \alpha_n)B + \alpha_n(1 - S) \qquad \text{if previous trial not rewarded}$$

Where α_r is a learning rate from rewarded trials, α_n is a learning rate from non-rewarded trials, and S is a variable which takes value 1 if the up second-step state was reached on the trial and 0 if the down second-step state was reached.

The variable B determined the value of both the second-step states (used only for predicting dopamine activity in Figures 4 & S6) and first-step actions as:

$$V_{Inf}(up) = B$$

$$V_{Inf}(down) = (1 - B)$$

$$Q_{Inf}(a \rightarrow up) = B$$

$$Q_{Inf}(a \rightarrow down) = (1 - B)$$

Where $V_{Inf}(up)$ and$V_{Inf}(down)$ are the values of the up and down second-step states, and $Q_{Inf}(a \rightarrow up)$ and $Q_{Inf}(a \rightarrow down)$ are the values of the actions which commonly lead to the up and down state.

<u>Reward-as-cue</u>

We considered another strategy in which the most recent trial's outcome acts as a discriminative stimulus for the state of the task, termed 'Reward-as-cue' in [11]. This strategy has fixed habits (here expressed as first-step action values) deployed depending on the previous trial's outcome and second-step state where it was obtained:

If previous trial rewarded:

$$Q_{RC}(a \rightarrow s) = z$$

$$Q_{RC}(a \rightarrow s') = 0$$

If previous trial not rewarded:

$$Q_{RC}(a \rightarrow s) = 0$$

$$Q_{RC}(a \rightarrow s') = (1 - z)$$

Where $Q_{RC}(a \rightarrow s)$ is the first-step action that commonly leads to the second-step state reached on the previous trial, $Q_{RC}(a \rightarrow s')$ is the first-step action that commonly leads to the second-step state not reached on the previous trial, and z is a variable between 0 and 1 which determines the relative strength with which reward and reward omission affect the next trial choice.

<u>Combined action values</u>:

A set of different candidate models were created by combining one of more of the above strategies in a weighted sum with (optionally) bias and perseveration parameters, to give net action values:

$$Q_{net}(a) = \sum_i w_i Q_i(a) + K(a)$$

Where w_i is the weight assigned to strategy i whose first step action values are given by $Q_i(a)$, and $K(a)$ is the modifier to the value of first-step action a due to any bias or perseveration terms included in the model. In models where bias was included, this increased the value of the left action by an amount determined by a bias strength parameter on all trials. In models where perseveration was included, this increased the value of the first step action chosen on the previous trial by an amount determined by a perseveration strength parameter.

The combined action values determined choice probabilities via a softmax decision rule as:

$$P(a) = \frac{e^{Q_{net}(a)}}{\sum_a e^{Q_{net}(a)}}$$

Model fitting and comparison

We generated a total of 84 different individual models from the following classes:

<u>Model-free (MF)</u>*:* These models used only the model-free strategy, and varied with respect to whether they had asymmetric learning rates for positive and negative outcomes, and whether they included perseveration and/or bias.

<u>Model-based (MB)</u>*:* These models used only the model-based strategy, and varied with respect to whether they had asymmetric learning rates for positive and negative outcomes, and whether they included perseveration and/or bias.

<u>Hybrid (MF + MB)</u>*:* These models used both the model-based and model-free strategy, and varied with respect to whether they had asymmetric learning rates for positive and negative outcomes, and whether they included perseveration and/or bias.

<u>Inference</u>*:* These models used the inference and/or the Reward-as-cue strategies, and varied with respect to whether they learned from both rewarded and non-rewarded outcomes or just rewarded outcomes, and whether they included perseveration and/or bias.

Each model was fit separately to data from each subject using maximum a posteriori fitting with priors: Beta (α=2, β=2) for unit range parameters, Gamma (α=2, β=0.4) for positive range parameters, and Normal (σ=5) for unconstrained parameters. The optimisation was repeated 15 times starting with randomised initial parameter values drawn from the prior distributions, and the best of these fits was used. Model comparison used the Bayesian Information criterion, which was computed per subject , and summed across subjects. In Figure 1G we show only the best fitting model from each of the 4 above classes, the full set of compared models is shown in Figure S1A.

The overall best fitting model combined the inference and Reward-as-cue learning strategies with bias and perseveration. Both strategies were only influenced by rewarded but not non-rewarded trials - i.e., in the inference strategy the learning rate for non-rewarded trials $\alpha_n = 0$, and in the Reward-as-cue strategy, the parameter determining asymmetry between rewarded and non-rewarded trials $z = 1$.

To compare data simulated from the model with real data (Figure S1), we simulated 520 sessions of 350 trials each, using parameters values drawn from the distribution of each animal's fits.

Photometry analysis

Photometry signals were aligned across trials by linearly time-warping the signal at the two points in the trial where timings were determined by subject behaviour; between initiation and choice, and between the second-step port illuminating and being poked (Figure 2B). Activity at other time periods was not warped.

For the analyses presented in Figures 3, 4, S3-6, we used linear regression to predict trial-by-trial dopamine activity at each timepoint in a trial as:

$$y(i,t) = \sum_p \beta_p(t) X_p(i) + \beta_o(t) + \varepsilon(i,\ t)$$

where $y(i,t)$ is the calcium z-scored activity on trial i at timepoint t, $\beta_p(t)$ is the weight for predictor p at timepoint t, $X_p(i)$ is the value of the predictor p on trial i, $\beta_o(t)$ is the intercept at timepoint t, and $\varepsilon(i,\ t)$ is the residual unexplained variance.

The linear regression was fit separately for each subject to obtain the coefficient time-courses $\beta_p(t)$. For each predictor we plot the mean and standard error across subjects. Statistical significance of each predictor at each timepoint was assessed using a t test comparing the distribution of coefficients across subjects with zero, with Benjamini-Hochberg correction for comparison of multiple timepoints.

The linear regressions in Figures 3 & S3-4 used the following predictors:

- *Reward*: +0.5 if current trial is rewarded, -0.5 otherwise
- *Good Second-Step*: +0.5 if the second step reached on the current trial has high reward probability, -0.5 if low reward probability, 0 if neutral block.
- *Correct Choice*: +0.5 if subject current trial choice commonly leads to the high reward probability second-step port, -0.5 if to the low reward probability second-step port, 0 if neutral block.
- *Repeat Choice*: +0.5 if same choice as previous choice, -0.5 if different choice to previous trial.
- *Direct Reinforcement Action Value Update*: +0.5/-0.5 if current choice is the same as the previous choice and the previous trial was rewarded/not-rewarded, 0 if different choice from previous trial.
- *Model-Consistent Action Value Update*: +0.5/-0.5 if current choice commonly leads to the previous second-step when it was rewarded/not-rewarded, -0.5/+0.5 if current choice rarely leads to the previous second-step when it was rewarded/not-rewarded.
- *Previous Reward, Same Second-Step*: +0.5 if current second-step is the same as in the previous trial and previous trial was rewarded, -0.5 if same second-step and previous trial was unrewarded, 0 if current second-step different from second-step on the previous trial.
- *Previous Reward, Different Second-Step*: +0.5 if current second-step is different from previous second-step and previous trial was rewarded, -0.5 if different second-step and previous second-step was unrewarded, 0 if current second-step is the same as the second-step on the previous trial.
- *Common Transition*: +0.5 if a common transition occurs on the current trial, -0.5 if a rare transition.
- *Reward Rate*: exponential moving average of the recent reward rate (tau=10 trials).
- *Contralateral Choice*: +0.5 if current choice is in the contralateral side from the recording site, -0.5 if current choice is in the ipsilateral side.
- *Up Second-Step:* +0.5 if current second-step is up, -0.5 if current second-step is down.

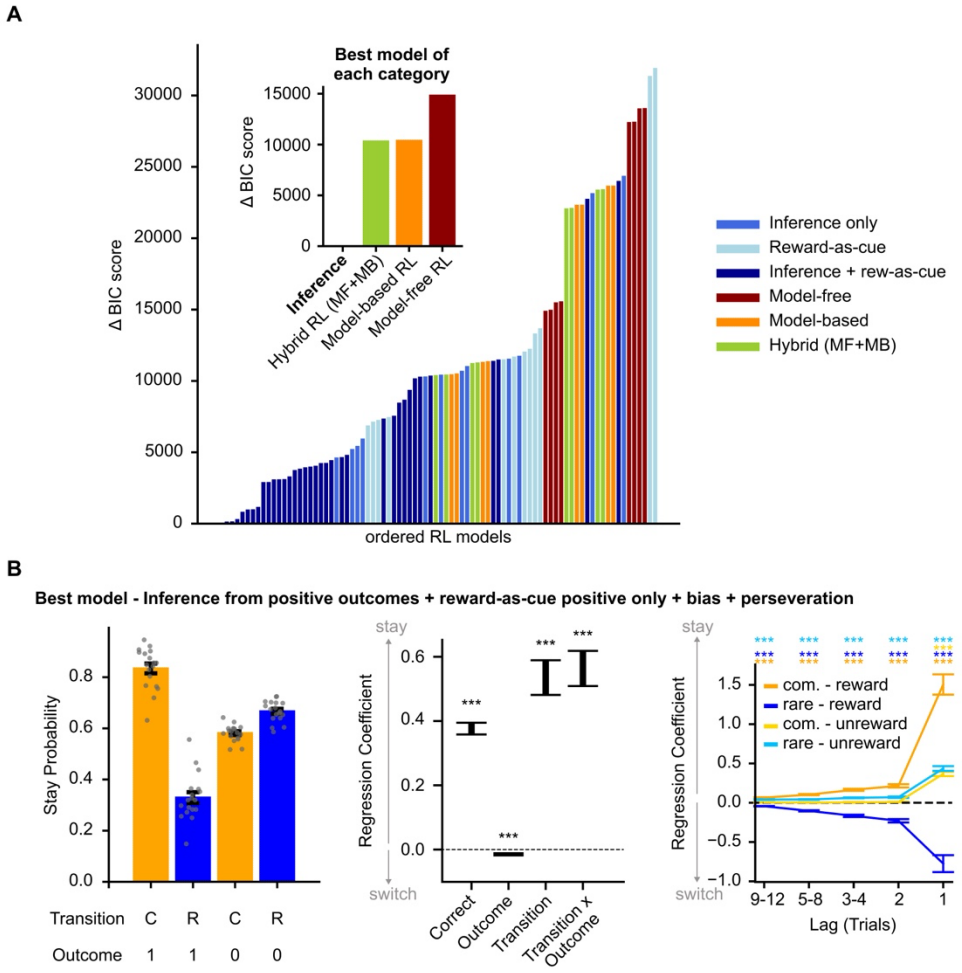
The linear regression in Figure S6 use the same predictors as above (Figures 3 & S4-5) except predictors *Previous Reward, Same Second-Step*, *Previous Reward, Different Second-Step* and *Model-Consistent Action Value Update*, which were replaced with predictors split by outcome as:

- *Same vs Different Second-Step, Previously Rewarded*: +0.5/-0.5 if current second-step is the same/different as in the previous trial and the previous trial was rewarded, 0 if the previous trial was not rewarded.
- *Same vs Different Second-Step, Previously Non-Rewarded*: +0.5/-0.5 if current second-step is the same/different as in the previous trial and the previous trial was not rewarded, 0 if the previous trial was rewarded.
- *Model-consistent Action Value Update from Rewarded Trials*: +0.5/-0.5 if current choice commonly/rarely leads to the previous second-step and the previous trial was rewarded, 0 if the previous trial was not rewarded.
- *Model-consistent Action Value Update from Unrewarded Trials*: -0.5/+0.5 if current choice commonly/rarely leads to the previous second-step and the previous trial was not rewarded, 0 if the previous trial was rewarded.

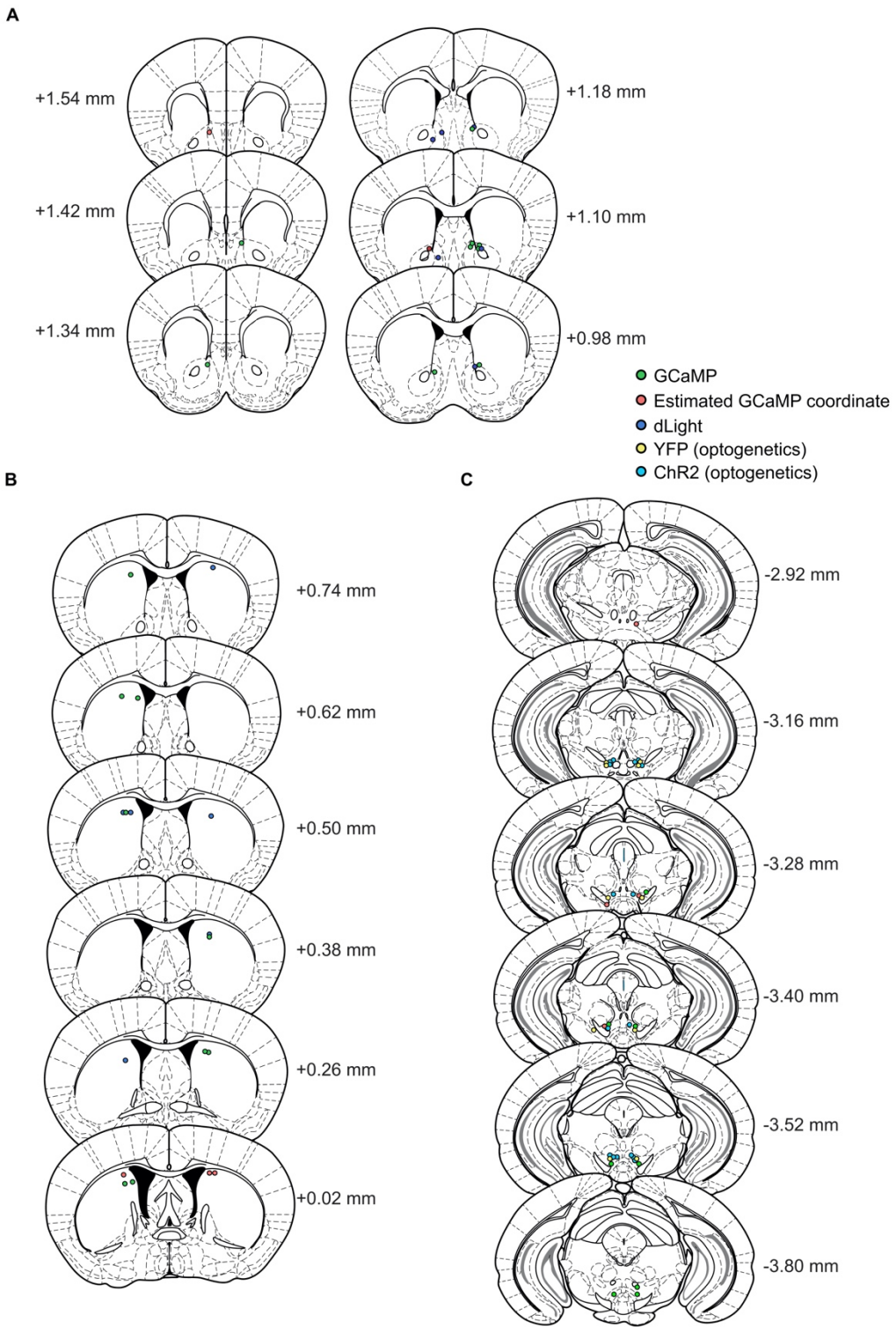
The linear regressions used in figures 4 & S6 used the above-described *Reward* & *Reward Rate* regressors with the following additional regressors:

- *Previous Reward:* +0.5 if previous trial was rewarded, -0.5 otherwise
- *Second-Step Value:* The value $V_{Inf}(s)$ of the second-step reached on the current trial from the best fitting RL model.
- *Chosen Action Value*: The value $Q_{Inf}(c)$ of the first-step action chosen on the current trial from the best fitting RL model.

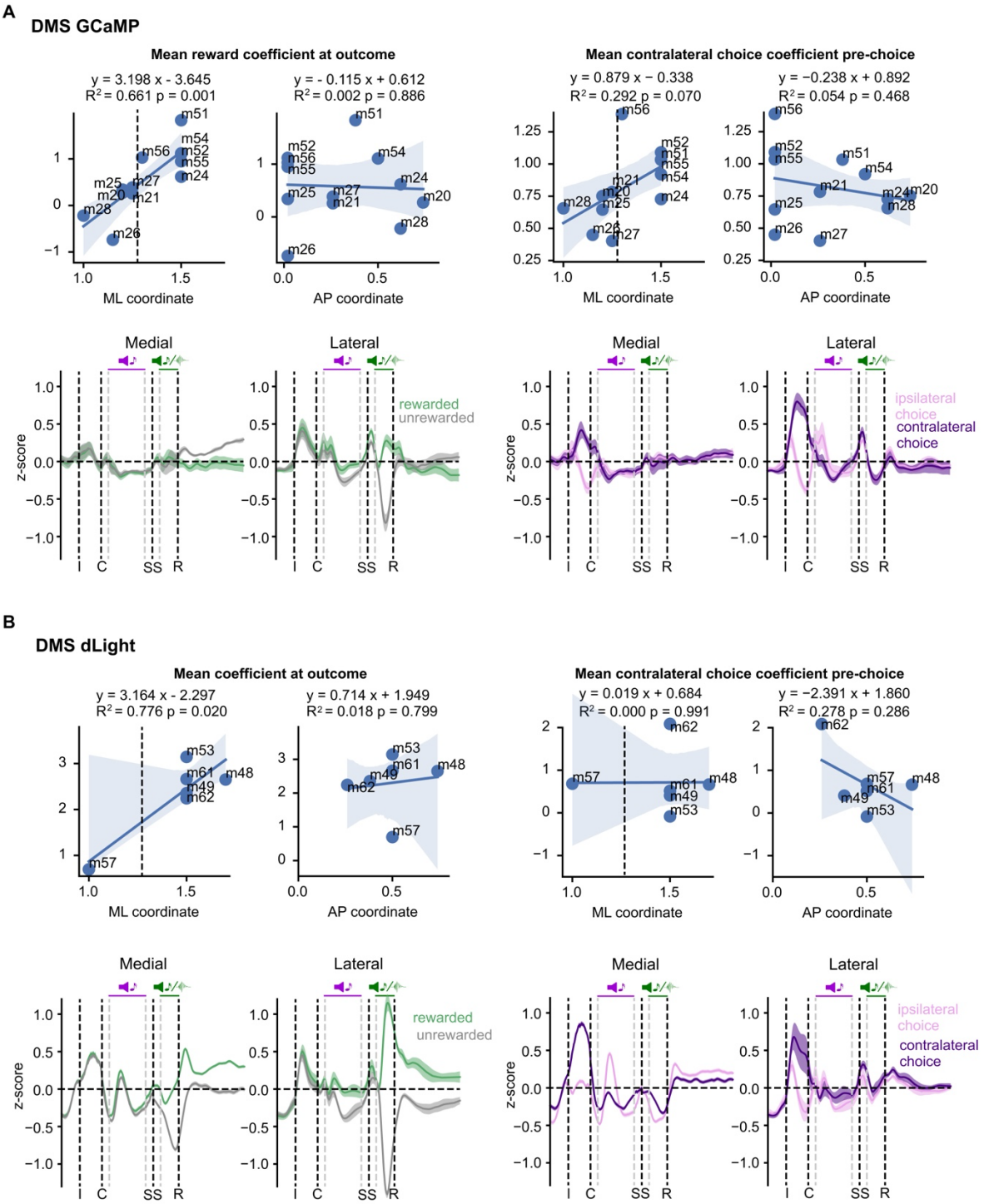
SUPPLEMENTARY FIGURES AND TABLES



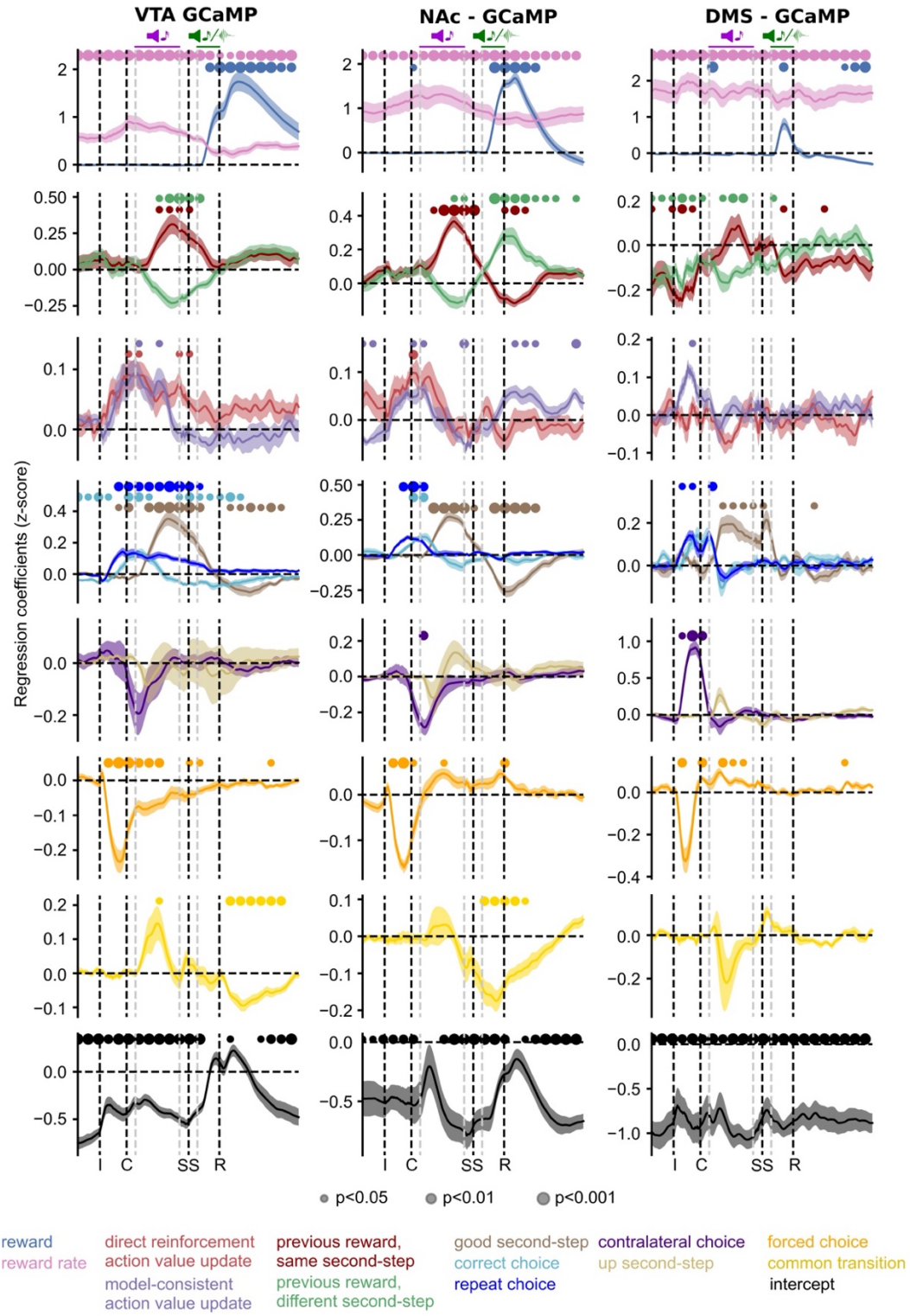
Supplementary Figure 1. Model comparison and simulated behaviour. A) BIC score model-comparison of all the models tested (see Supplementary Table 2 for details of each model). Colours indicate the model's strategy. Each strategy was tested both with and without choice bias and perseveration. Additionally, models were tested with both symmetric and asymmetric learning rates for rewarded and unrewarded trials, and for models using 'Inference' and 'Reward-as-cue' strategies also with the influence of unrewarded outcomes set to 0 rather than parameterised. The best fitting model labelled 'Inference' in Figure 1G and the inset panel combines both 'Inference' *and* 'Reward-as-cue' strategies (see main text and methods). In (A) models that use either 'Inference' or 'Reward-as-cue' strategies, or both, are differentiated by shade of blue. Fitted parameter values for the best model in each class are presented in Supplementary Table 1. **B)** Stay probability and logistic regression analysis of data simulated for the best fitting model, which combined 'Inference' and 'Reward-as-cue' strategies (both only using rewarded outcomes), choice bias and perseveration. *p<0.05, **p<0.01, ***p<0.001.



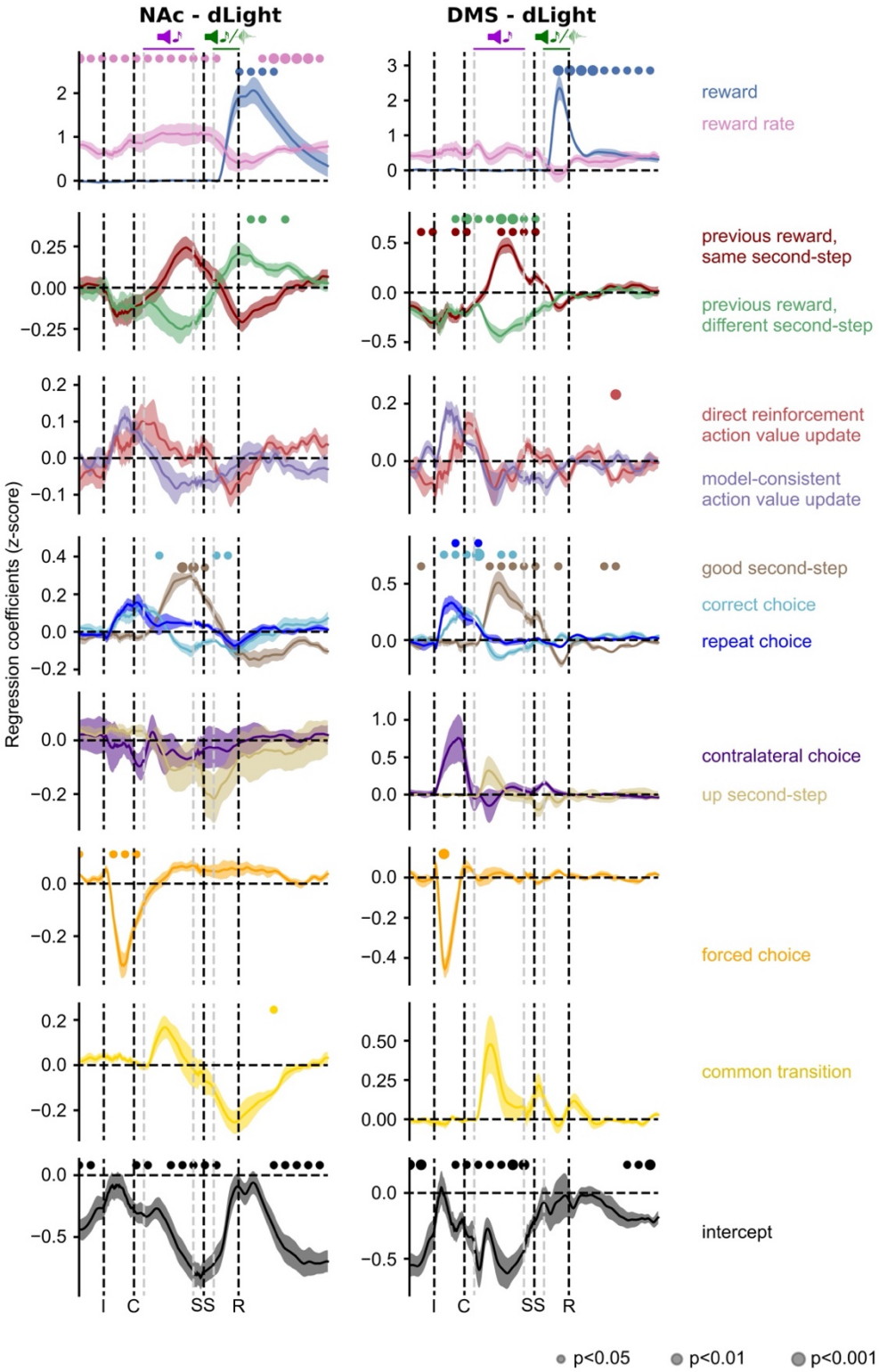
Supplementary Figure 2. Fibre placement in **A)** NAc, **B)** DMS, and **C)** VTA. Green, GCaMP6f; red, GCaMP6f animals where the tip of the fibre placement is estimated; dark blue, dLight1.1; yellow, YFP animals from the optogenetics experiment; light blue, ChR2 animals.



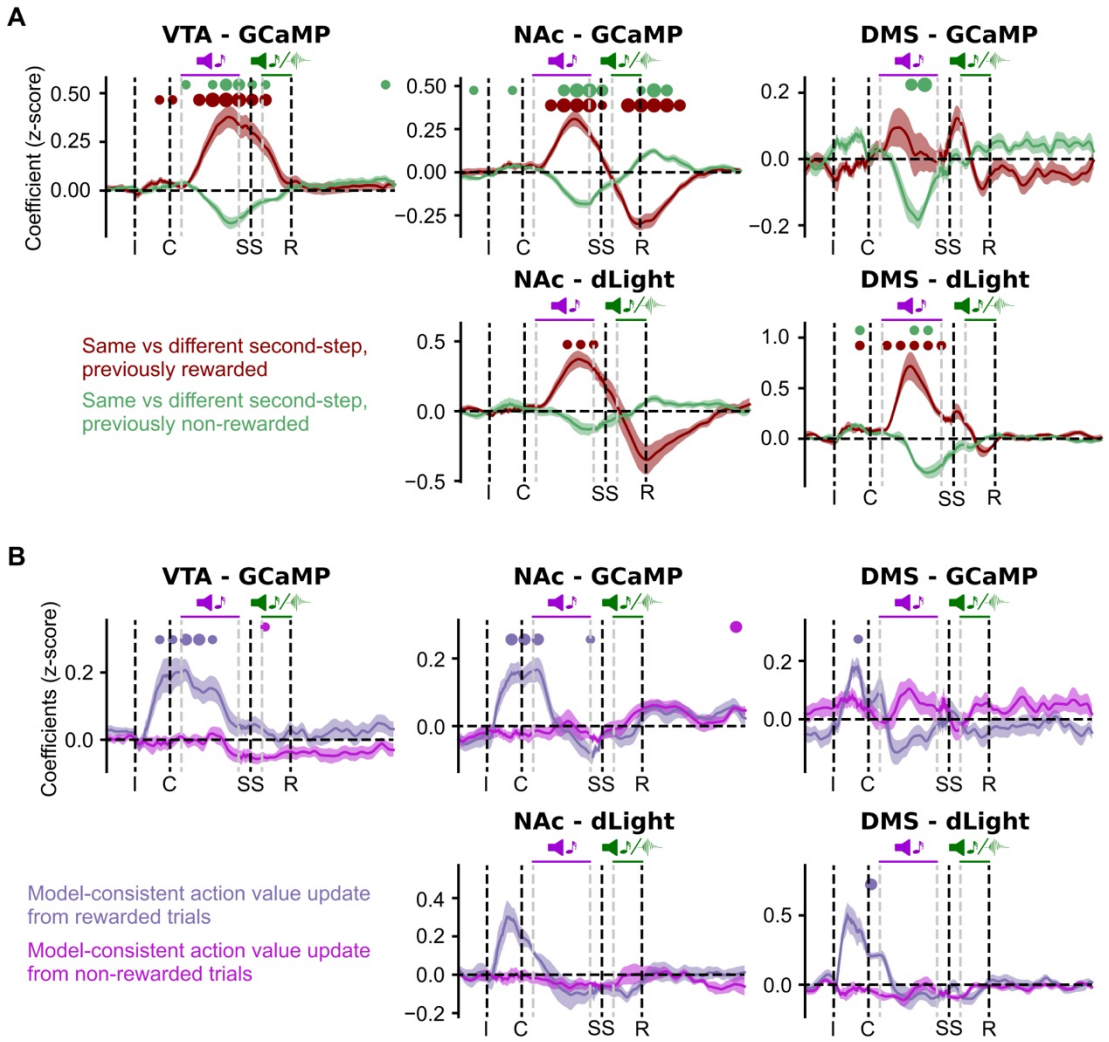
Supplementary Figure 3. Medio-lateral gradient of reward responses in DMS. A) Dopamine activity measured through GCaMP. **B)** Dopamine receptor binding measured through dLight. *Left*, mean reward coefficient (from the linear regression model) at the time of outcome cue; *Right*, mean contralateral choice coefficient pre-choice. *Top*, correlation between mean coefficient and medio-lateral (ML) or antero-posterior (AP) location of the optic fibre. *Bottom*, mean z-scored activity split by outcome or lateralised choice for animals with medial or lateral placement in DMS. Dashed line in the correlation plots show how animals were divided in two groups (medial and lateral).



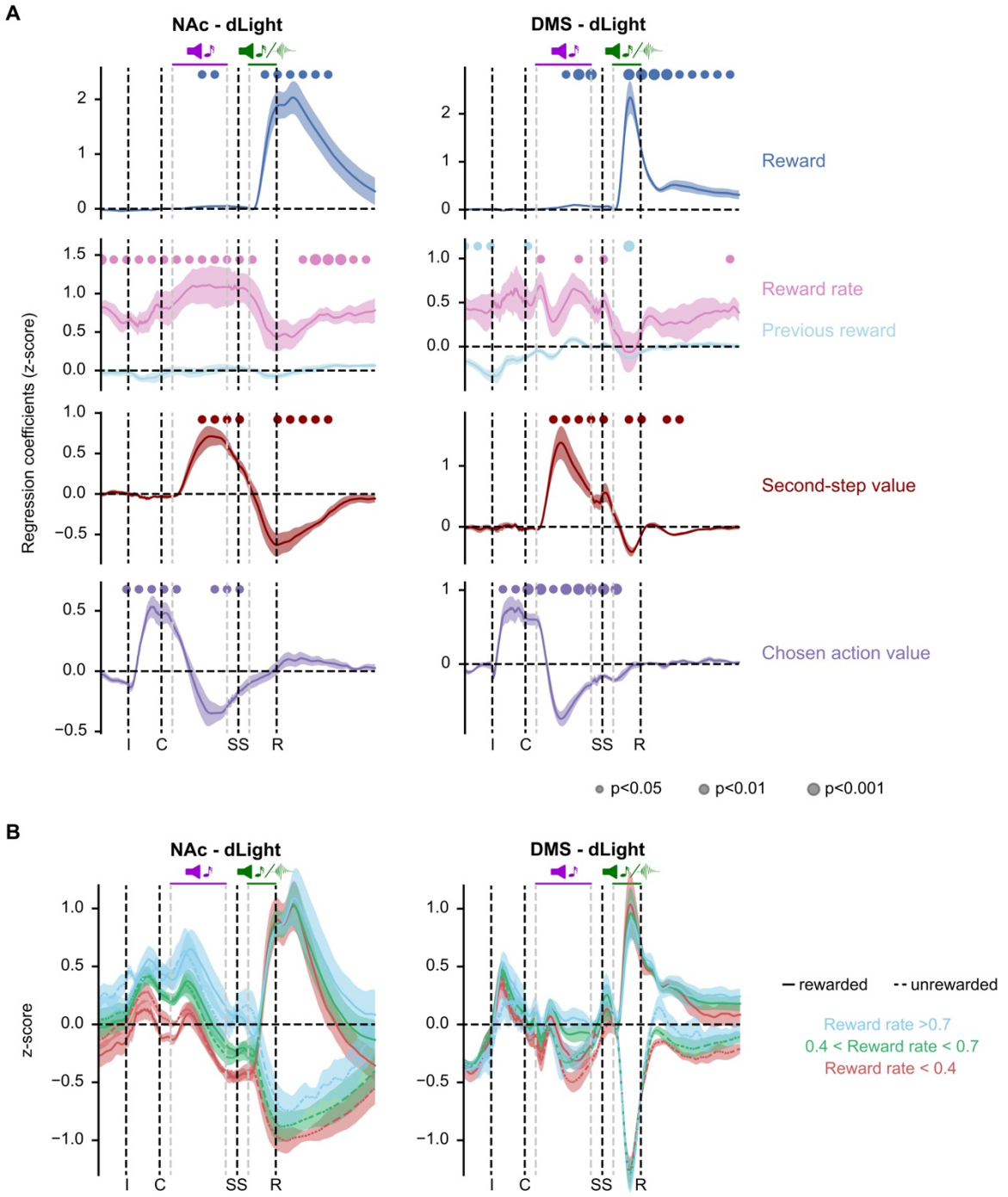
Supplementary Figure 4. Regression coefficients from all the predictors used in the behavioural logistic regression model predicting VTA, NAc and DMS dopamine activity (GCaMP recordings). Shaded area indicates cross subject standard error. Shaded area indicates cross subject standard error. Dots represents statistical significance, two-sided t-test comparing the cross-subject distribution against 0, after Benjamini-Hochberg multiple comparison correction, small dot: $p < 0.05$, medium dot: $p < 0.01$, large dot: $p < 0.001$.



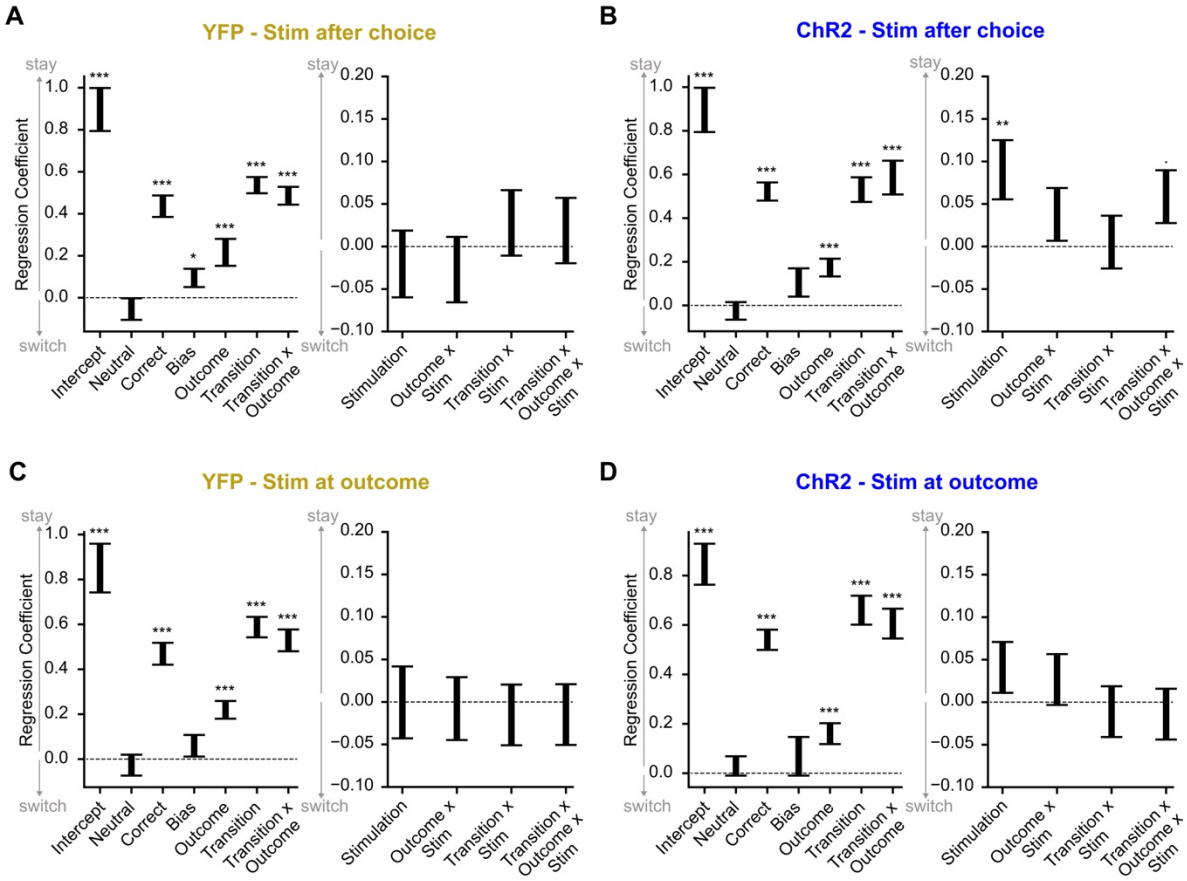
Supplementary Figure 5. Regression coefficients from all the predictors used in the behavioural logistic regression model predicting NAc and DMS dopamine receptor binding (dLight recordings). Shaded area indicates cross subject standard error. Dots represents statistical significance, two-sided t-test comparing the cross-subject distribution against 0, after Benjamini-Hochberg multiple comparison correction, small dot: $p < 0.05$, medium dot: $p < 0.01$, large dot: $p < 0.001$.



Supplementary Figure 6. Value update regressors split by rewarded/not rewarded outcome. A) Second-step value update based on whether previous trial was rewarded or not. **B)** Model-based action value update split by outcome on the previous trial. Shaded area indicates cross subject standard error. Dots represents statistical significance, two-sided t-test comparing the cross-subject distribution against 0, after Benjamini-Hochberg multiple comparison correction, small dot: $p < 0.05$, medium dot: $p < 0.01$, large dot: $p < 0.001$.



Supplementary Figure 7. A) Regression model using second-step and action values derived from the best RL fitting model predicting dopamine concentrations (dLight) in NAc (left) and DMS (right). Shaded area indicates cross subject standard error. Dots represents statistical significance, two-sided t-test comparing the cross-subject distribution against 0, after Benjamini-Hochberg multiple comparison correction, small dot: $p < 0.05$, medium dot: $p < 0.01$, large dot: $p < 0.001$. **B)** Mean z-scored signal across a trial split between rewarded (continuous trace) and unrewarded (dashed trace) trials, and split between high reward rate (>0.7 exponential moving average of rewards, tau=10) in blue, medium reward rate (between 0.4 and 0.7) in green, and low reward rate (<0.4) in red.



Supplementary Figure 8. Mixed effects logistic regression model predicting repeat choice after optogenetic stimulation in the YFP controls (**A, C**) and ChR2 group (**B, D**) on sessions when stimulation occurred after choice (**A, B**) or at outcome (**C, D**). •p<0.1, *p<0.05, **p<0.01, ***p<0.001.

		Fitted parameters							
Best fitting model from each class	**BIC**	**MF weight**	**MB weight**	**Infer. weight**	**rew-as-cue weight pos. out.**	**learning rate**	**eligibility trace**	**bias**	**persv.**
Inference from positive outcomes + reward-as-cue (pos. only) + bias + persv.	143574.73			1.55 ± 0.44	1.76 ± 0.86	0.29 ± 0.05		0.02 ± 0.21	0.75 ± 0.23
Hybrid (MF+MB) + bias + persv.	153983.96	0.59 ± 0.37	1.66 ± 0.61			0.46 ± 0.18	0.26 ± 0.19	0.01 ± 0.21	1.31 ± 0.36
MB + bias + persv.	154043.15		2.18 ± 0.58			0.40 ± 0.16		0.01 ± 0.21	1.31 ± 0.35
MF + bias + persv	158494.59	1.93 ± 0.46				0.33 ± 0.11	0.48 ± 0.12	0.02 ± 0.18	1.35 ± 0.37

998
999 **Supplementary Table 1. Best RL model fits**. Summed BIC score and fitted parameters for each
1000 model (Mean ± standard deviation across subjects) (Models shown in Figure 1G and inset
1001 subplot in Figure S1). Best model is the inference strategy learning only from positive
1002 outcome, with added reward-as-cue only from positive outcomes, bias and perseveration.
1003

			Learning rate			Reward-as-cue			Kernels	
Rank	Model Category	BIC	Sym.	Asym.	Positive outcome only	Sym.	Asym.	Positive outcome only	Bias	Persev.
1	Inference + Reward-as-cue	143574.7			■			■	■	■
2	Inference + Reward-as-cue	143721.2		■				■	■	■
3	Inference + Reward-as-cue	143737.4			■		■		■	■
4	Inference + Reward-as-cue	143898.1		■			■		■	■
5	Inference + Reward-as-cue	144408.8			■			■		■
6	Inference + Reward-as-cue	144568.3		■				■		■
7	Inference + Reward-as-cue	144580.7			■		■			■
8	Inference + Reward-as-cue	144751.1		■			■			■
9	Inference + Reward-as-cue	146491.8	■					■	■	■
10	Inference + Reward-as-cue	146497.7			■			■	■	
11	Inference + Reward-as-cue	146680.1	■				■		■	■
12	Inference + Reward-as-cue	146686.8		■				■	■	
13	Inference + Reward-as-cue	146697.8			■		■		■	
14	Inference + Reward-as-cue	146886.4		■			■		■	
15	Inference + Reward-as-cue	147333.0	■					■		■
16	Inference + Reward-as-cue	147425.5			■	■			■	■
17	Inference + Reward-as-cue	147528.0	■				■			■
18	Inference + Reward-as-cue	147573.9		■		■			■	■
19	Inference + Reward-as-cue	147634.3			■			■		
20	Inference + Reward-as-cue	147826.0		■				■		
21	Inference + Reward-as-cue	147834.6			■		■			
22	Inference + Reward-as-cue	148025.9		■			■			
23	Inference	148217.0		■					■	■
24	Inference + Reward-as-cue	148237.9			■	■				■
25	Inference + Reward-as-cue	148394.8		■		■				■
26	Inference	148808.4			■				■	■
27	Inference	149025.5		■						■
28	Inference	149540.3			■					■
29	Reward-as-cue	150458.0	■				■		■	■

30	Reward-as-cue	150727.1								
31	Reward-as-cue	150844.0								
32	Inference + Reward-as-cue	150937.0								
33	Reward-as-cue	151047.0								
34	Inference + Reward-as-cue	151137.2								
35	Inference + Reward-as-cue	152058.5								
36	Inference + Reward-as-cue	152258.8								
37	Inference + Reward-as-cue	152948.5								
38	Inference + Reward-as-cue	153757.6								
39	Inference + Reward-as-cue	153876.3								
40	Inference	153888.9								
41	Inference + Reward-as-cue	153952.6								
42	Hybrid (MF+MB)	153984.0								
43	Inference	154019.7								
44	Hybrid (MF+MB)	154027.2								
45	Model-based	154043.2								
46	Model-based	154102.6								
47	Inference	154287.3								
48	Inference	154622.8								
49	Hybrid (MF+MB)	154832.4								
50	Hybrid (MF+MB)	154872.3								
51	Model-based	154909.2								
52	Model-based	154961.6								
53	Inference + Reward-as-cue	154983.4								
54	Inference + Reward-as-cue	155079.1								
55	Reward-as-cue	155089.5								
56	Inference	155134.9								
57	Reward-as-cue	155281.3								
58	Inference	155339.3								
59	Reward-as-cue	155634.4								
60	Reward-as-cue	155830.7								
61	Reward-as-cue	156907.8								
62	Reward-as-cue	157269.1								

63	Model-free	158494.6	■						■	■
64	Model-free	158565.4		■					■	■
65	Model-free	159081.5	■							■
66	Model-free	159155.1		■						■
67	Hybrid (MF+MB)	165671.7		■					■	
68	Hybrid (MF+MB)	165703.5	■						■	
69	Model-based	165921.0	■						■	
70	Model-based	165923.4		■					■	
71	Inference + Reward-as-cue	166337.3	■			■			■	
72	Inference	166734.5	■						■	
73	Hybrid (MF+MB)	166982.4		■						
74	Hybrid (MF+MB)	167012.9	■							
75	Model-based	167258.9		■						
76	Model-based	167264.1	■							
77	Inference + Reward-as-cue	167595.2	■			■				
78	Inference	167948.4	■							
79	Model-free	171716.9	■						■	
80	Model-free	171741.7		■					■	
81	Model-free	172670.7	■							
82	Model-free	172698.3		■						
83	Reward-as-cue	174946.6	■			■			■	
84	Reward-as-cue	175508.4	■			■				

Supplementary Table 2. Model comparison. Table indicates features included in each model considered in the model comparison. The models varied with respect to the learning strategies they employed (indicated by *Model Category*). Additionally, models were tested with both *Symmetrical (*"Sym.*")* and *Asymmetrical (*"Asym.*")* learning rates for rewarded and unrewarded trials, and for models using the 'inference' strategy, also with the influence of unrewarded outcomes set to 0 rather than parameterised (*Positive outcome only)*. The symmetric or asymmetrical influence of rewarded and unrewarded trials was also tested in the Reward-as-cue strategy. Models were also tested with and without *Choice Bias* and/or *Perseveration* ("Persev.").