Mesolimbic dopamine encodes reward prediction errors independent of learning rates

Andrew Mah,[1] Carla Golden,[1] Christine Constantinople[1*]

[1]Center for Neural Science, New York University,

[*]To whom correspondence should be addressed; E-mail: constantinople@nyu.edu.

Summary

Biological accounts of reinforcement learning posit that dopamine encodes reward prediction errors (RPEs), which are multiplied by a learning rate to update state or action values. These values are thought to be represented in synaptic weights in the striatum, and updated by dopamine-dependent plasticity, suggesting that dopamine release might reflect the product of the learning rate and RPE. Here, we leveraged the fact that animals learn faster in volatile environments to characterize dopamine encoding of learning rates. We trained rats on a task with semi-observable states offering different rewards, and rats adjusted how quickly they initiated trials across states using RPEs. Computational modeling and behavioral analyses showed that learning rates were higher following state transitions, and scaled with trial-by-trial changes in beliefs about hidden states, approximating normative Bayesian strategies. Notably, dopamine release in the nucleus accumbens encoded RPEs independent of learning rates, suggesting that dopamine-independent mechanisms instantiate dynamic learning rates.

Introduction

Reinforcement learning describes how animals or agents learn the value of states and actions to select actions that maximize future expected rewards[1]. Standard reinforcement learning algorithms, including temporal-difference learning, update state and action values using reward prediction errors (RPEs), or the difference between experienced and expected rewards. The rate of error-driven learning is often assumed to be constant, but studies across species, including humans, monkeys, rats, and mice, have found behavioral evidence for dynamic learning rates[2–7]. In volatile environments, dynamic learning rates allow animals to learn faster when the world is changing, and more slowly when the world is stable[8–10].

A major target of midbrain dopamine neurons is the striatum, the input structure of the basal ganglia. In reinforcement learning accounts of the basal ganglia, inputs to the striatum, including from the cortex, are thought to convey the state the animal is, with the state or state-action values stored in the strength of those corticostriatal synaptic weights[1,11,12], whose plasticity is dopamine-dependent[13–16]. Thus, cortical inputs corresponding to high state values (i.e., stronger synapses) are more likely to drive activity in the striatum for action selection[12]. In this framework, corticostriatial synapses are strengthened or weakened proportional to dopamine RPEs. However, reinforcement learning algorithms update state values (or state-action values) proportional to the *product* of the RPE and learning rate. It is currently unclear whether dopamine conveys only the RPE, with other substrates dictating the learning rate, or whether dopamine encodes the RPE scaled by the learning rate (i.e., their product). These scenarios are indistinguishable if learning rates are static. Here, we leveraged the fact that reinforcement learning is dynamic in changing environments to characterize dopamine encoding of learning rates by recording dopamine release in rats performing a task with latent reward states.

We trained rats on a self-paced temporal wagering task in which they waited variable and un-

predictable amounts of time for explicitly cued water rewards. We introduced semi-observable states with blocks of trials that differed in their reward statistics. Rats adjusted how quickly they initiated trials based on the value of the state, consistent with a temporal-difference learning algorithm[17]. Notably, initiation times provided a continuous behavioral read-out of rats' estimates of state values, facilitating analysis of the dynamics with which they were updated over trials. Computational modeling and behavioral analyses revealed striking behavioral signatures of dynamic learning rates following block transitions. Learning rates appeared to scale with trial-by-trial changes in beliefs about hidden states, and behavioral data was inconsistent with alternative formulations of dynamic learning rates that scaled with salience[18] or model-free estimates of volatility or state uncertainty (i.e., Pearce-Hall models[19]). We show that scaling learning rates with changes in beliefs about hidden states approximates normative Bayesian changepoint detection strategies[20]. Finally, we measured dopamine release in the nucleus accumbens using fiber photometry of GRAB_{DA} sensors while rats performed the task. We found that dopamine release did not reflect learning rates, but did reflect beliefs about whether the rat was likely to be in a rewarded or unrewarded trial as rats waited for uncertain rewards[21,22]. These results show that beliefs about hidden states dictate rats' learning rates, and that dopamine encodes RPEs independent of learning rates.

Results

Rats use a dynamic learning rate.

We trained rats on a self-paced temporal wagering task with semi-observable reward blocks[17] (Fig. 1A-B). Rats were offered different volumes of water rewards (5, 10, 20, 40, 80 μL), cued by an auditory tone. On 75-85% of trials, rewards were delivered after variable, unpredictable delays drawn from an exponential distribution, and on 15-25% of trials, rewards were withheld. The rats could choose to wait during the delay for the water reward, or could opt-out at any

time to start a new trial. We introduced uncued blocks of trials with differing reward statistics: low blocks, which offered the smallest three rewards (5, 10, 20 μL), and high blocks, which offered the largest three rewards (20, 40, 80 μL), interleaved with mixed blocks, which offered all rewards (Fig 1B).

We measured the time between the rat's final poke in the reward or opt-out port and the start of the next trial (trial initiation time). Trial initiation times were inversely proportional to the value of the environment, and provided a continuous behavioral readout of rats' estimates of state values[17]. Rats were slower to initiate trials in low blocks compared to high blocks (Fig. 1C-D). Furthermore, when we regressed initiation times against previous reward offers, more recent offers had larger coefficients, and coefficients gradually decreased for more distant trials (Fig. 1E). This pattern is consistent with canonical reinforcement learning algorithms which estimate the value of the environment as a recency-weighted average of previous rewards. In previous work, we found that rats use state inference when deciding how long to wait, but initiate trials based on state value estimates that are computed with standard reinforcement learning algorithms[17]. Therefore, we focused on trial initiation times (as opposed to wait times) for this study.

In reinforcement learning, the decay time of previous trial coefficients is directly proportional to the learning rate parameter, which is often assumed to be static. However, examination of trial initiation times aligned to transitions from low or high blocks into mixed blocks suggests two phases of learning- an initial phase of fast learning, followed by slower dynamics later in the block, suggestive of higher learning rates immediately following block transitions (Fig. 2A). Previous work found that the initiation time transition dynamics could not be explained by a static learning rate[17]. Consistent with this result, when we regressed trial initiation times against previous rewards separately for the first and last ten trials of each mixed block, rats integrated over fewer trials earlier in the block, that is, had higher learning rates, compared to later

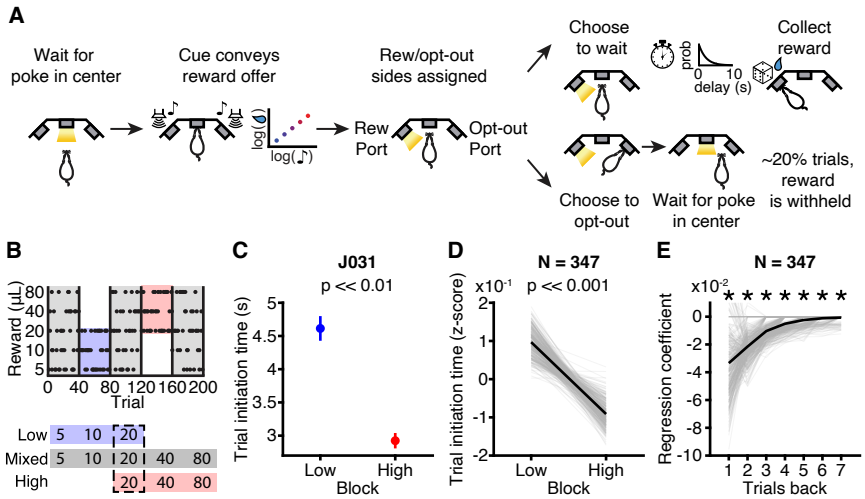


Figure 1: **Trial initiation times are sensitive to previous rewards and blocks A.** Task schematic. **B.** Block structure. **C-D.** (C) Average trial initiation times for a single rat ($p < 0.01$, Wilcoxon rank-sum test) and (D) across the population ($p << 0.001$, Wilcoxon signed-rank test, $N = 347$) are sensitive to block. **E.** Regression coefficients of previous rewards predicting trial initiation times across rats (Asterisk indicate $p < 0.05$, Wilcoxon signed-rank test, $N = 347$). All error bars are mean $\pm$ S.E.M.

in the block (Fig. 2B-C). Across rats, exponential functions fit to the regression coefficients had significantly smaller time-constants (higher learning rates) for early trials compared to late trials (Fig. 2D).

We next fit a simple reinforcement learning model with a static learning rate separately to early and late trials (Fig. 2E). The model estimated the value of the environment according to the recursive equation: $V_{t+1} = V_t + \alpha(R_t - V_t)$, where $R_t - V_t$ is the reward prediction error (RPE), and α is a learning rate that dictates the dynamics with which values are updated over trials. The trial initiation times were modeled as inversely proportional to this value, which captured the rat's behavior on held-out test data[17] (Fig. 2F). Models fit to early mixed block trials had significantly higher learning rates than models fit to late mixed block trials (Fig. 2G). Model comparison (negative log-likelihood per trial) confirmed that learning rates fit to early and late parts of the block provided better fits to held out test data in early and late block trials,

respectively (Fig. 2H). Overall, these data suggest that rats use a dynamic learning rate that is higher following block transitions compared to later in the block.

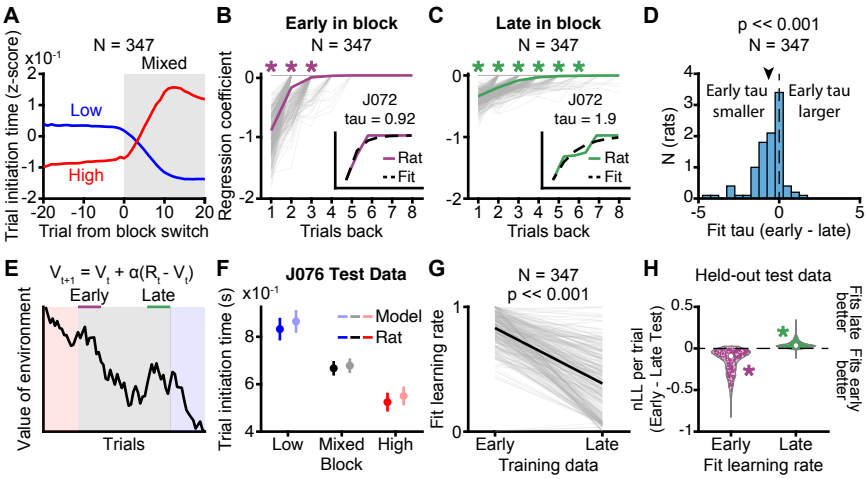


Figure 2: **Rats use a higher learning rate early in mixed blocks A.** Trial initiation times aligned to transitions into mixed blocks from low (blue) and high (red) blocks. Smoothed with a causal filter of size 10 trials. **B-C.** Previous reward regression coefficients for (B) early mixed block trials and (C) late mixed block trials (Wilcoxon sign-rank test, N = 347). Insets show example rat regression coefficient (solid line) overlaid with exponential fit (dashed line) **D.** Difference in time-constant, tau, of exponential decay fit to regression coefficients from early and late mixed block trials across rats. Arrow indicates mean ($p < 0.001$, paired Wilcoxon Sign-rank test, N = 347). **E.** Model schematic. **F.** Example rat behavior and model prediction on held-out late trial test data. **G.** Recovered learning rate parameters for early and late mixed block training data across rats ($p < 0.001$, Wilcoxon Sign-rank test, N = 347) **H.** Model comparison of early and late parameters on held-out test data. White circles indicate mean. (Wilcoxon sign-rank test, N = 347). Asterisks indicate $p < 0.05$. Arror bars are mean $\pm$ S.E.M.

Rats' dynamic learning rates reflect changing beliefs.

Previous work has shown that animals adjust their learning rates depending on the volatility in the environment, since it is advantageous to learn faster in dynamic environments[3,5,7]. To determine which features of the environment the rats were using to adjust their learning rates, we tested several classes of dynamic learning rate models (Fig. 3A). Previous studies have proposed that subjects might scale their learning rates based on salient events or outcomes[18,23].

We modeled salience as a linear function of log-reward offer on each trial, such that larger rewards were assumed to be more salient than smaller rewards. We refer to this salience-based learning rate as the Mackintosh model, following [18].

Alternatively, previous work suggests that it is advantageous to modulate learning rates based on perceived volatility in the environment[3,5,24]. Specifically, when volatility is high, the world might be changing, so agents should exhibit higher learning rates to disregard outcomes further in the past, which are unlikely to be predictive of future states. We tested two classes of volatility-based dynamic learning rate models. First, the Pearce-Hall model adjusts the learning rate proportional to the unsigned reward prediction error (RPE; experienced - expected reward) of the previous trial[19]. Intuitively, large RPEs indicate that reward expectations are inaccurate, potentially because the environment is changing. Second, we used a model we previously developed in [17], which we refer to as the ΔBelief model. This model uses Bayes' rule to infer the probability of being in each block given the rat's current reward offer and past reward history. The learning rate scales with the trial-by-trial change in the posterior belief about the block (Methods). If beliefs are changing dramatically across trials, it is likely that the environment is changing.

We used these models to generate qualitative predictions of how initiation times might vary over trials. For instance, the models make distinct predictions about how the variance of the initiation times should vary within mixed blocks. Because the mixed blocks include all reward offers, the magnitude of RPEs are comparable both early and late in the blocks. Therefore, the Pearce-Hall model, which scales the learning rate by the previous unsigned RPE, predicts similar trial-by-trial learning (i.e., initiation time variance) early and late in the block. Similarly, large rewards are equally likely early and late in mixed blocks, so the Mackintosh model, which scales the learning rate with the log-reward offer, predicts similar initiation time variance on the early and late trials. In contrast, the ΔBelief model, which scales the learning rate with the

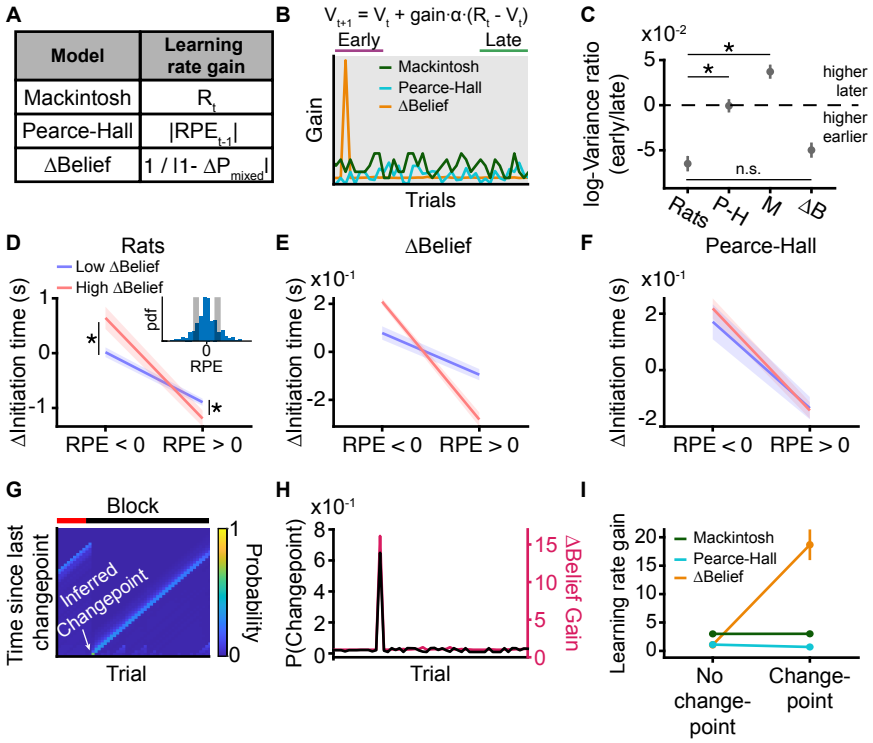


Figure 3: **Rats use changing beliefs to modulate learning rates A.** Dynamic learning rate models. **B.** Examples of learning rate gain from different dynamic learning rate models. **C**. Trial initiation time variance early vs. late for dynamic learning rate models and rat data. (Wilcoxon rank-sum test, N = 347). **D.** Change in trial initiation time for negative and positive RPE bins (inset) for large (red) or small (blue) change in mixed block belief ($>$ or $<$ 50th percentile; Wilcoxon sign-rank test, N = 347). **E-F**. Model predictions for (E) ΔBelief model and (F) Pearce-Hall model. **G.** Simulated changepoint posterior for Bayesian Online Changepoint Detection model around a block transition. Upper bar shows true block. **H.** ΔBelief model learning rate gain correlates with changepoint probability from Bayesian Online Changepoint Detection model. **I**. Learning rate gain for dynamic learning rate models on inferred changepoint trials from the Bayesian Online Changepoint Detection model. Asterisks indicate $p < 0.05$. Error bars are mean $\pm$ S.E.M.

trial-by-trial change in the belief about the inferred state, predicts higher variance in initiation times early in the block, compared to late when beliefs are stable (Fig. 3B). We measured the variance of the initiation times for the first and last ten trials of each mixed block, and computed their log ratio (negative values correspond to larger trial-by-trial changes in behavior early in the block). Consistent with the ΔBelief model, but not the other models, the variance of rats'

initiation times was higher in the early block trials, compared to later trials with stable beliefs (Fig. 3C).

Next, we explicitly tested whether rats exhibited more trial-by-trial learning following small or large changes in posterior beliefs about the block, controlling for RPEs. We focused on early trials in the block, when there should be a broader range of beliefs to increase statistical power. We estimated RPEs using model parameters fit to behavior from the first ten trials of each block. We compared the trial-by-trial change in initiation times for the same values of RPEs, but conditioned on when the ΔBelief on the previous trial was high or low (less than or greater than 50th percentiles). We found that rats changed their initiation times more for both positive and negative RPEs following large ΔBelief trials, compared to trials with the same binned RPEs but low ΔBeliefs (Fig. 3D), which is consistent with the ΔBelief model (Fig. 3E), but not the other dynamic learning rate models (Fig. 3F for Pearce Hall; Supp. Fig. 1 for Mackintosh).

Finally, to provide a normative foundation for the ΔBelief model, we show that it approximates Bayesian online changepoint detection[20]. In our task, changepoint detection models would infer changes in the underlying distribution from which rewards are drawn; these inferred state transitions are referred to as changepoints. Specifically, the model calculates a posterior distribution over potential run-lengths, or the time since the last changepoint (Fig. 3G). However, changepoint detection is often computationally expensive, especially for sessions with many trials. On each trial, the number of potential run-lengths increases by one, which increases the number of computations the model performs. These computations are often simplified using mathematical tricks (e.g., conjugate priors[20]) which do not exist for discrete reward distributions, further increasing computational complexity for our task. We wondered if this strategy could be approximated by any of the dynamic learning rate models. Strikingly, we found that the gain term from the ΔBelief model was highest at inferred changepoints, suggesting a close correspondence between estimates of changing beliefs from these two models

(Fig. 3H). Over many simulated sessions, the gain term from the ΔBelief model was significantly higher at trials with inferred changepoints compared to other trials. By contrast, there was no systematic relationship between the unsigned RPE (Pearce-Hall model), or log reward offer (Mackintosh model), and inferred changepoints (Fig. 3I). This suggests that the ΔBelief model approximates normative changepoint detection strategies.

Dopamine encodes RPE independent of learning rate.

Dopamine release in the nucleus accumbens core (NAcc) is thought to act as an RPE for reinforcement learning[25–38]. Moreover, the NAcc is thought to determine the vigor of motivated behaviors[39], and we recently found that dopamine RPEs in the NAcc causally determine initiation times on subsequent trials[40]. Dopamine RPEs are thought to mediate plasticity at synapses onto medium spiny neurons, the principal cells of the striatum, such that animals will be more (or less) likely to take actions in a state that produced positive (or negative) RPEs. In this scenario, dopamine release in the striatum may represent the product of the learning rate and RPE, and should reflect dynamic learning rates in our task. Consistent with this hypothesis, NAcc dopamine release can in principle be modulated by beliefs about hidden states[21,22].

To test this hypothesis, we measured dopamine release using fiber photometery and $GRAB_{DA}$, a fluorescent dopamine sensor (Fig 4A). We observed robust phasic dopamine responses at the time of the offer cue that were consistent with an RPE. First, NAcc dopamine release scaled monotonically with offered reward volume in mixed blocks, with the dips for smaller rewards (Fig. 4B,D), consistent with a correlation between reward and RPE. An RPE signal should also scale inversely with expectations. Focusing on 20 μL trials, which appeared in all blocks, dopamine increased on 20 μL trials in low blocks, when reward expectations were low, and decreased on 20 μL trials in high blocks, when expectations were high (Fig. 4C,E). Third, when we regressed dopamine release (area under the curve from 0 to 0.5 seconds) against reward

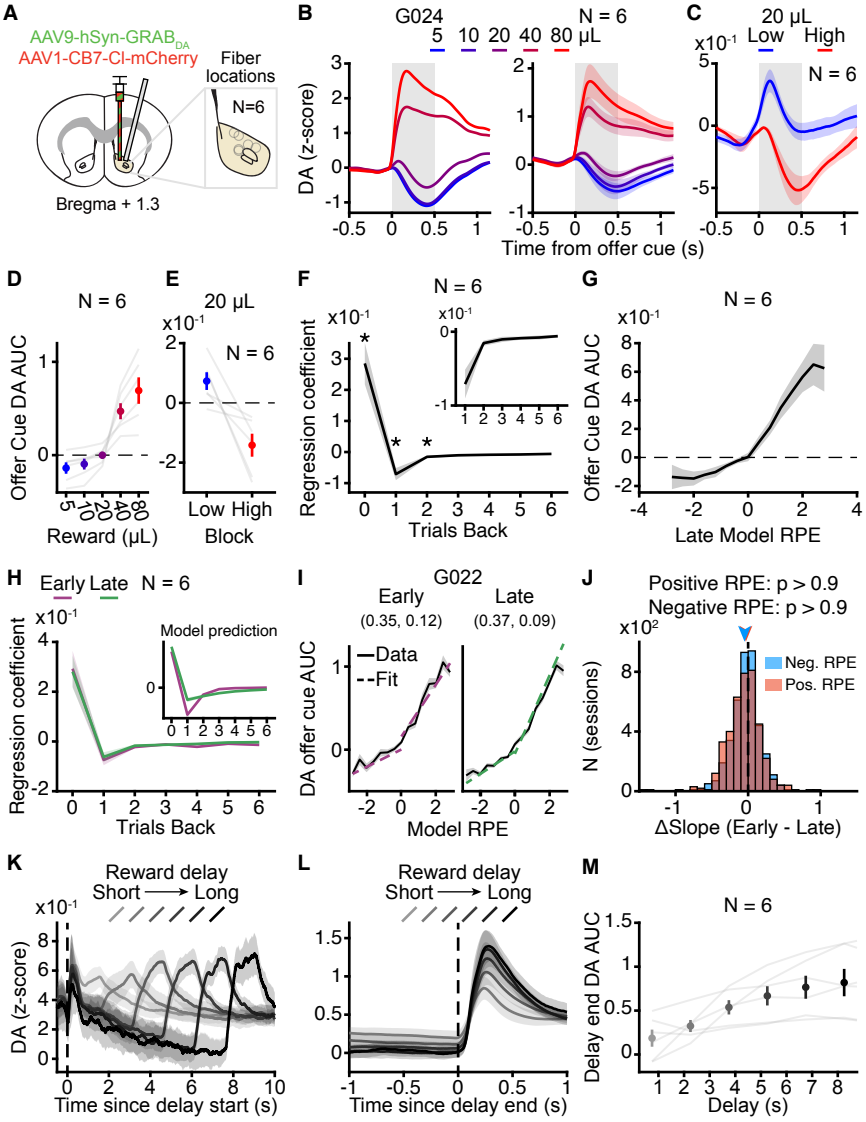


Figure 4: **Dopamine encodes reward prediction errors independently of learning rates A.** Recording sites. **B.** Dopamine response aligned to offer cue split by rewards in mixed blocks for an example rat (left) and averaged across rats (right; $N = 6$). **C.** Average dopamine response to offer cue for 20 μL split by block ($N = 6$). **D.** Average AUC from B. (gray box, N = 6). **E.** Average AUC from C. (gray box, N = 6). **F.** Reward history regression coefficients for dopamine signal. Inset shows previous reward coefficients (Sign-rank test, $N = 6$). **G.** Offer cue AUC by binned late model RPE ($N = 6$). **H.** Reward history regression coefficients for dopamine signals in early (purple) and late (green) trials. $p > 0.05$ for all coefficients, paired sign-rank test, $N = 6$. Inset shows model prediction for parameters fit to early vs. late trials. **I.** Example rat dopamine

Figure 4: (cont). response by binned RPE for early (left) and late (right) trials (solid). with positive and negative RPE regressions (dashed). Parentheticals indicate positive and negative slopes, respectively. **J.** Difference between early and late regression slopes for DA for positive (orange) and negative RPEs (blue) over individual sessions for all rats. Arrows indicate mean (N = 409 sessions, $p_{\text{early}}, p_{\text{late}} > 0.9$, one-tailed Wilcoxon sign-rank test. **K.** Dopamine response aligned to beginning of delay period for rewarded mixed block trials split by reward delay. **L.** Dopamine response aligned to end of delay period for rewarded mixed block trials as a function of reward delay. **M.** Baseline-corrected average AUC from L. (N-6). Asterisks indicate $p < 0.05$. Error bars are mean $\pm$ S.E.M.

history, we found positive coefficients for the current trial and negative coefficients for previous trials (Fig. 4F). Finally, dopamine release correlated with RPEs estimated from the behavioral model (Fig. 4F). In previous studies, it was technically challenging to fit the dynamic learning rate model[17], so we fit the static learning rate model to the last 10 trials of each block, and related dopamine release to RPEs on those trials only (Fig. 4G).

Next, we compared dopamine release early vs. late in the block when the rats used different learning rates. If dopamine release is the product of the reward prediction error and the learning rate, then dopamine should be influenced by fewer trials in the past when the learning rate is higher (Fig. 4H insert). This hypothesis also predicts that dopamine release would be greater for the same RPE earlier in the block, when the learning rate is higher, compared to later in the block, when beliefs are stable and learning rates are lower. However, when we compared dopamine release early and late in blocks, we found no significant differences in the magnitude of reward history regression coefficients (Fig. 4H) or the encoding of RPEs (Fig. 4I). To account for nonlinear encoding of RPEs, we fit separate regression lines to dopamine encoding (AUC) of positive and negative model RPEs for each session. There was no significant difference in the slope parameters fit to early or late trials in each session across rats, and this was true for both positive and negative RPEs (Fig. 4I,J; Supp. Fig. 2). Therefore, despite strong behavioral evidence for higher learning rates earlier in mixed blocks when beliefs about hidden states are changing, we did not find differential dopamine dynamics between early and late block trials.

Previous work found that dopamine release can reflect beliefs about hidden task states if rewards are probabilistic and variable in their timing[21]. To determine if dopamine exhibited similar dynamics in our task, we examined activity during the delay period. Because rewards were omitted on a subset of trials, as animals waited for the reward, there is ambiguity about whether they are in a rewarded or unrewarded trial. Sorting rewarded trials based on delay duration revealed negative dopamine ramps for the duration of the delay (Fig. 4K), consistent with previous studies[21]. These negative ramps were also apparent as different baselines when the data were aligned to the cue indicating that reward was available (Fig. 4L). In a temporal difference framework, these ramps could reflect moment-by-moment negative RPEs as the animals are waiting[32]. When a cue indicated that reward was available, there was a large phasic dopamine response, the magnitude of which scaled with the delay (Fig. 4K-M). This pattern has been interpreted to reflect beliefs about hidden states[21]: the probability of being in a catch trial increases over time, so the reward cue is more unexpected given beliefs about the state. Therefore, mesolimbic dopamine reflects beliefs about hidden states at the time of probabilistic rewards, but not about latent reward states at the time of the offer cue.

Discussion

We found that rats adjust their learning rates based on trial-by-trial changes in posterior beliefs about inferred states, but dopamine encodes RPEs independent of these changing beliefs. Previous work found that dopamine release can reflect beliefs about hidden states[21,22], for instance if rewards are probabilistic and variable in their timing[21]. In a conditioning paradigm with probabilistic rewards, during the delay period, there is ambiguity about the state the animal is in: whether it is in the delay period, or if it has transitioned to the inter-trial interval on an unrewarded trial. In other words, there is *uncertainty* about the current state. However, this type of uncertainty is qualitatively different from the uncertainty about the latent reward blocks in our

task. Uncertainty in the timing of probabilistic rewards reflects *expected*, or irreducible, uncertainty[24]. Even with a perfect model of the world, there is inherent uncertainty in the outcome of stochastic events - one can know the probabilities of a coin toss, but cannot predict the outcome of the next flip. This is in contrast to *unexpected* uncertainty, which is related to the volatility of the environment - a perfect model of the world now may not be perfect if the world changes[24,41]. In other words, unexpected uncertainty about the reward blocks can be reduced with additional observations, while expected uncertainty about probabilistic reward timing is irreducible. We find that dopamine release in the NAcc reflects expected but not unexpected uncertainty. Together, these results suggest that different types of uncertainty may map onto neurobiologically distinct mechanisms, with potentially dissociable consequences for learning[24,34,41–44].

Previous work has suggested that dopamine release in the NAcc directly encodes the learning rate[45], implicating dopamine in a broader class of policy learning algorithms, beyond traditional value learning. However, our work examines distinct phases of learning from [45]. We exclusively recorded from expert, as opposed to naive, animals. Presumably our expert rats have already learned their final behavioral policy, and must only learn about the current state of the environment. On the other hand, task naive animals are simultaneously learning the associative structure of the environment and how to optimally behave in that environment. Such distinct learning goals likely engage dopamine differently and could explain the differences in our findings.

One key future question is what could be driving the dynamic learning rate at the level of synaptic plasticity. Previous work has shown that plasticity at corticostriatial synapses depends on the coordinated activity of dopamine[13–16] with other neuromodulators, like acetylcholine[46] and serotonin[47–49]. In particular, serotonin neurons, which project to the NAcc from the dorsal raphe nucleus[50], have been shown to encode unexpected uncertainty in a probabilistic reward task with uncued blocks of trials in mice[7]. Other neuromodulators that have been hypothe-

sized to encode unexpected uncertainty, like norepinephrine[41], do not strongly project to the NAcc[51–53], but could influence learning in other neural circuits. Future studies clarifying task-related dynamics of other neuromodulators, as well as dynamics of distinct classes of glutamatergic inputs to the striatum, may elucidate the circuit mechanisms that combine dopaminergic reward prediction errors with trial-by-trial changes in beliefs to modulate the rate of learning at behavioral and synaptic levels.

References

1. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction* (MIT press, 2018).
2. Nassar, M. R., Wilson, R. C., Heasly, B. & Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience* **30,** 12366–12378 (2010).
3. Behrens, T. E., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. Learning the value of information in an uncertain world. *Nature neuroscience* **10,** 1214–1221 (2007).
4. Hayden, B. Y., Heilbronner, S. R., Pearson, J. M. & Platt, M. L. Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *Journal of Neuroscience* **31,** 4178–4187 (2011).
5. Nassar, M. R. *et al.* Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience* **15,** 1040–1046 (2012).
6. Funamizu, A., Ito, M., Doya, K., Kanzaki, R. & Takahashi, H. Uncertainty in action-value estimation affects both action choice and learning rate of the choice behaviors of rats. *European Journal of Neuroscience* **35,** 1180–1189 (2012).
7. Grossman, C. D., Bari, B. A. & Cohen, J. Y. Serotonin neurons modulate learning rate through uncertainty. *Current Biology* **32,** 586–599 (2022).
8. Amari, S. A theory of adaptive pattern classifiers. *IEEE Transactions on Electronic Computers,* 299–307 (1967).
9. Sutton, R. S. *Adapting bias by gradient descent: An incremental version of delta-bar-delta* in *AAAI* **92** (1992), 171–176.
10. Murata, N., Kawanabe, M., Ziehe, A., Müller, K.-R. & Amari, S.-i. On-line learning in changing environments with applications in supervised and unsupervised learning. *Neural Networks* **15,** 743–760 (2002).
11. Joel, D., Niv, Y. & Ruppin, E. Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural networks* **15,** 535–547 (2002).
12. Doya, K. Reinforcement learning: Computational theory and biological mechanisms. *HFSP journal* **1,** 30 (2007).
13. Centonze, D. *et al.* Unilateral dopamine denervation blocks corticostriatal LTP. *Journal of neurophysiology* **82,** 3575–3579 (1999).
14. Kerr, J. N. & Wickens, J. Dopamine D-1/D-5 receptor activation is required for long-term potentiation in the rat neostriatum in vitro. *Journal of neurophysiology* **85,** 117–124 (2001).
15. Reynolds, J. N., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413,** 67–70 (2001).

16. Shen, W., Flajolet, M., Greengard, P. & Surmeier, D. J. Dichotomous dopaminergic control of striatal synaptic plasticity. *Science* **321,** 848–851 (2008).

17. Mah, A., Schiereck, S. S., Bossio, V. & Constantinople, C. M. Distinct value computations support rapid sequential decisions. *Nature communications* **14,** 7573 (2023).

18. Mackintosh, N. J. A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological review* **82,** 276 (1975).

19. Pearce, J. M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review* **87,** 532 (1980).

20. Adams, R. P. & MacKay, D. J. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742* (2007).

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

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

23. Iigaya, K. Adaptive learning and decision-making under uncertainty by metaplastic synapses guided by a surprise detection system. *Elife* **5,** e18073 (2016).

24. Soltani, A. & Izquierdo, A. Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience* **20,** 635–644 (2019).

25. Olds, J. Self-stimulation of the brain: Its use to study local effects of hunger, sex, and drugs. *Science* **127,** 315–324 (1958).

26. Corbett, D. & Wise, R. A. Intracranial self-stimulation in relation to the ascending dopaminergic systems of the midbrain: a moveable electrode mapping study. *Brain research* **185,** 1–15 (1980).

27. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275,** 1593–1599 (1997).

28. Waelti, P., Dickinson, A. & Schultz, W. Dopamine responses comply with basic assumptions of formal learning theory. *Nature* **412,** 43–48 (2001).

29. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47,** 129–141 (2005).

30. Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B. & Uchida, N. Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *nature* **482,** 85–88 (2012).

31. Day, J. J., Roitman, M. F., Wightman, R. M. & Carelli, R. M. Associative learning mediates dynamic shifts in dopamine signaling in the nucleus accumbens. *Nature neuroscience* **10,** 1020–1028 (2007).

32. Kim, H. R. *et al.* A unified framework for dopamine signals across timescales. *Cell* **183,** 1600–1616 (2020).

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

34. Parker, N. F. *et al.* Reward and choice encoding in terminals of midbrain dopamine neurons depends on striatal target. *Nature neuroscience* **19,** 845–854 (2016).

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

36. Tsai, H.-C. *et al.* Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science* **324,** 1080–1084 (2009).

37. Hamid, A. A. *et al.* Mesolimbic dopamine signals the value of work. *Nature neuroscience* **19,** 117–126 (2016).

38. Adamantidis, A. R. *et al.* Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *Journal of neuroscience* **31,** 10829–10835 (2011).

39. Floresco, S. B. The nucleus accumbens: an interface between cognition, emotion, and action. *Annual review of psychology* **66,** 25–52 (2015).

40. Golden, C. E. *et al.* Estrogenic control of reward prediction errors and reinforcement learning. *bioRxiv,* 2023–12 (2023).

41. Angela, J. Y. & Dayan, P. Uncertainty, neuromodulation, and attention. *Neuron* **46,** 681–692 (2005).

42. O'reilly, J. X. Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in neuroscience* **7,** 33773 (2013).

43. Payzan-LeNestour, E. & Bossaerts, P. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology* **7,** e1001048 (2011).

44. McGuire, J. T., Nassar, M. R., Gold, J. I. & Kable, J. W. Functionally dissociable influences on learning rate in a dynamic environment. *Neuron* **84,** 870–881 (2014).

45. Coddington, L. T., Lindo, S. E. & Dudman, J. T. Mesolimbic dopamine adapts the rate of learning from action. *Nature* **614,** 294–302 (2023).

46. Reynolds, J. N. *et al.* Coincidence of cholinergic pauses, dopaminergic activation and depolarisation of spiny projection neurons drives synaptic plasticity in the striatum. *Nature Communications* **13,** 1296 (2022).

47. Burke, D. A. & Alvarez, V. A. Serotonin receptors contribute to dopamine depression of lateral inhibition in the nucleus accumbens. *Cell reports* **39** (2022).

48. Pommer, S., Akamine, Y., Schiffmann, S. N., de Kerchove d'Exaerde, A. & Wickens, J. R. The effect of serotonin receptor 5-HT1B on lateral inhibition between spiny projection neurons in the mouse striatum. *Journal of Neuroscience* **41,** 7831–7847 (2021).

49. Mathur, B. N., Capik, N. A., Alvarez, V. A. & Lovinger, D. M. Serotonin induces long-term depression at corticostriatal synapses. *Journal of Neuroscience* **31,** 7402–7411 (2011).

50. Van Bockstaele, E. J., Biswas, A. & Pickel, V. M. Topography of serotonin neurons in the dorsal raphe nucleus that send axon collaterals to the rat prefrontal cortex and nucleus accumbens. *Brain research* **624,** 188–198 (1993).

51. Allin, R, Russell, V., Lamm, M. & Taljaard, J. Regional distribution of monoamines in the nucleus accumbens of the rat. *Neurochemical research* **13,** 937–942 (1988).

52. Delfs, J. M., Zhu, Y., Druhan, J. P. & Aston-Jones, G. S. Origin of noradrenergic afferents to the shell subregion of the nucleus accumbens: anterograde and retrograde tract-tracing studies in the rat. *Brain research* **806,** 127–140 (1998).

53. McKittrick, C. R. & Abercrombie, E. D. Catecholamine mapping within nucleus accumbens: differences in basal and amphetamine-stimulated efflux of norepinephrine and dopamine in shell and core. *Journal of neurochemistry* **100,** 1247–1256 (2007).

54. Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191,** 507–520 (2007).

55. Creamer, M. S., Chen, K. S., Leifer, A. M. & Pillow, J. W. Correcting motion induced fluorescence artifacts in two-channel neural imaging. *PLoS computational biology* **18,** e1010421 (2022).

Methods

Subjects

A total of 347 Long-evans rats (*Rattus norvegicus*, 215 males, 134 females) between the ages of 6 and 24 months old. This cohort included 24 TH-Cre rats, 8 ADORA2A-Cre, and 3 DRD1-Cre rats. Animals use procedures were approved by the New York University Animal Welfare Committee (UAWC #2021-1120) and carried out in accordance with National Institute of Health standards.

Rats were typically pair-housed. To motivate behavioral performance, rats were water restricted from Monday to Friday, during which time they received water during behavioral training sessions, typically 90 minutes, followed by 20 minutes of ad libitum water. Rats were given ad libitum water following training on Friday though mid-day Sunday. Rats were weighed daily.

Behavioral training

We have previously published a detailed description of the behavioral shaping procedure for this task[17]. Rats performed a self-paced temporal wagering task. Rats initiated trials by maintaining a nose poke in the center port for a variable period drawn from a uniform distribution over [0.8, 1.2] seconds. As the rat maintained the nose poke, the reward offer on that trial was conveyed by an auditory tone [1, 2, 4, 8, 16 kHz], which mapped onto one of five rewards ([5, 10, 20, 40, 80 μL] for males, [2, 4, 8, 16, 32 μL] for females). Following the reward tone presentation, rats could either wait a random delay drawn from an exponential distribution with mean of 2.5 seconds to receive their reward, or could opt-out at any time to immediately start a new trial. On 15-25% of trials (catch trials), reward was withheld to force the rats to exercise the opt-out option.

Training for male and female rats

We collected data from both male and female rats (215 males, 134 females). Male and female rats were trained with the same shaping procedure. Early cohorts of female rats experienced the same reward set as the males. However, because female rats are smaller, they consumed less water and performed substantially fewer trials than the males. Therefore, to obtain sufficient behavioral trials from both, females reward offers were slightly reduced while maintaining the logarithmic spacing: [4, 8, 16, 32, 64 μL]. For behavioral analysis, reward volumes were treated as equivalent to the corresponding volume for the male rats (e.g., 16 μL trials for female rats were treated the same as 20 μL trials for male rats). The auditory tones were identical to those used for male rats. We did not observe any significant differences between the male and female rats, in terms of contextual effects, or behavioral dynamics at block transitions[17]. Photometry data in this study was collected from females.

Criteria for including behavioral data

To determine whether rats sufficiently understood the mapping between auditory cues and water reward volumes, we evaluated their wait times on catch trials as a function of offered rewards. For each session, we first removed wait times that were greater than two standard deviations from the mean, which likely reflected lapses in attention/task disengagement. Next, we regressed wait time against offered reward. We included sessions with significant positive slopes that preceded at least one other session with a positive slope. We excluded trials with trial initiation times above the 99th percentile of the rat's cumulative trial initiation time distribution pooled over sessions.

Behavioral modeling

To model trial initiation times, we developed computational models based on [17, 54], which describe the optimal trial initiation time, TI, given the value of the environment, V, as

$$TI = \frac{D}{V}$$

where D is a scale parameter. We developed multiple computational models that instantiated different algorithms for estimating the value of the environment. We estimate the value of the environment on trial t, V_t, by the recursive formula

$$V_{t+1} = V_t + \alpha_t(R_t - V_t)$$

where $\alpha_t = g_t \cdot \alpha_0$ is the learning rate, g is the learning rate gain, and R_t is the $\log_2$(reward) on the current trial. For the static learning rate model, $g = 1$ for all trials.

Dynamic learning rate models

We tested several models of dynamic learning rates.

1. **Mackintosh surprise model**

In this model, the gain on the learning rate is proportional to the salience of that trial[18], which we assumed to be directly proportional to the reward offer volume on that trials so,

$$g_t = \log_2(R_t)$$

where α_0 is the base learning rate.

2. **Pearce-Hall model**

In this model, the learning rate gain is directly proportional to the inferred volatility of the environment. Volatility in this model is "model-free" as estimated as the unsigned RPE on the

previous trial[19], so

$$g_t = |\text{RPE}_{t-1}|$$

where RPE_{t-1} is the reward prediction error on the previous trial.

3. Δ **Belief model** In this model, as in 2, the learning rate is directly proportional to the inferred volatility of the environment. In this model, volatility is calculated using the trial-by-trial change in the belief of being in a mixed block, using Bayes rule and knowledge of the underlying block structure[17], so

$$g_t = \frac{1}{1 - |B_{t-1} - B_t|}$$

where $B_t = P(\text{Block} = \text{Mixed} \,|\, R_t)$. We used the mixed block probability as a summary statistic for the full posterior distribution over blocks, as there is always some ambiguity about whether the animal is in a mixed block, and the block probabilities all need to sum to one. Therefore, changes in the probability of being in a mixed block reflect changes in the full posterior distribution on each trial[17].

Fitting and evaluating models

We fit the models by minimizing the negative-log-likelihood of the the model using MATLAB's constrained minimization function, *fmincon*, assuming log-normal noise with constant variance (variance = 1.7, selected from cross-validated grid search on a subset of rats). We used 100 random seeds and selected the fit with the lowest negative-log-likelihood. We have previously validated our fitting procedure by fitting the models to generative datasets with known parameters[17]. We used 5-fold cross-validation to fit five sets of parameters to each rat (one for each fold), and selected the parameters with the lowest negative log-likelihood per trial on that fold's test set. Finally, we evaluated the performance of the model fits on a final held-out validation set of trials.

We fit the static learning rate model to the rats' trial initiation times in early and late trials separately. From previous work, sequential learning effects were primarily driven by post-violation trials[17], so we fit the model to only post-violation trials. Furthermore, the distribution of trial initiation times was generally heavy-tailed, and seemed to reflect multiple processes on different interacting timescales (e.g., reward sensitivity on short timescales, attention, motivation, and satiety on longer timescales). To capture only task-engaged trials, we removed trial initiation times above the 90th percentile of trial initiation times pooled over sessions for each rat.

Bayesian Online Changepoint Detection model

We compared the dynamic learning rate models to a normative Bayesian online changepoint detection model[20]. This model identifies abrupt changes, or changepoints, in the underlying generative distribution of sequentially observed data, which in our case corresponds to block transitions. The time between changepoints is called the run-length. At a high level, on trial t, the model looks at the last N trials (N ranges from 0 to t) and estimates the probability that these N observations come from a different distribution than the trials before them. If that probability is high, then the model returns a run-length of N, meaning a changepoint occurred N trials ago.

Let x_t denote the observation on trial t and $\mathbf{x}_{t_1:t_2}$ be the sequence of observations from t_1 to t_2, inclusive, i.e., $\{x_{t_1}, x_{t_1+1}, ..., x_{t_2-1}, x_{t_2}\}$. On trial t, the run-length, r_t can range from 0 to t. Finally, given a run-length, r_t, let $\mathbf{x}_t^{(r_t)}$ be the observations since the last changepoint, that is, $\mathbf{x}_{t-r_t:t}$. We calculate the probability of each potential run-length, known as the run-length posterior, with

$$P(r_t \mid \mathbf{x}_{1:t}) = P(r_t, \mathbf{x}_{1:t})/P(\mathbf{x}_{1:t})$$

We can simplify the above by marginalizing over the previous run-lengths, r_{t-1}, applying the

chain rule, and the assumptions that our data are independently generated, giving us

$$P(r_t, x_{1:t}) = \sum_{r_{t-1}} P(r_t \mid r_{t-1}) P(x_t \mid r_{t-1}, \mathbf{x}_t^{(r_{t-1})}) P(r_{t-1}, \mathbf{x}_{1:t-1})$$

The first term, $P(r_t \mid r_{t-1})$ is called the changepoint prior and captures how often changepoints occur, which depends on the hazard rate. Given a previous run-length, r_{t-1}, the next run-length can only be $r_{t-1} + 1$ (a changepoint did not occur) or 0 (a changepoint did occur). As described in [17], for simplicity, we assume that the hazard rate is constant with a value of $1/40$, so we have

$$P(r_t \mid r_{t-1}) = \begin{cases} \frac{1}{40}, & \text{if } r_t = 0 \\ 1 - \frac{1}{40}, & \text{if } r_t = r_{t-1} + 1 \\ 0, & \text{else} \end{cases}$$

The second term, $P(x_t | r_{t-1}, \mathbf{x}_t^{(r_{t-1})})$, is called the predictive probability. This term calculates the high level intuition given above: given some hypothetical run-length, are the data since that run-length consistently from one distribution. To calculate this, we assume that the rats have knowledge of the underlying block structure. We can calculate the predictive probability by marginalizing over the blocks, B, giving us

$$P(x_t \mid r_{t-1}, \mathbf{x}_t^{(r_{t-1})}) = \sum_B P(x_t \mid B) P(B \mid r_{t-1}, \mathbf{x}_t^{(r_{t-1})})$$

The first term is simply the likelihood of x_t given a block. We can use Bayes rule to calculate the second term, giving us

$$\begin{aligned} P(B \mid r_{t-1}, \mathbf{x}_t^{(r_{t-1})}) &\propto P(\mathbf{x}_t^{(r_{t-1})} \mid B) P(B) \\ &= \prod_{l=t-r_{t-1}}^{t} P(x_l \mid B) P(B) \end{aligned}$$

which calculates the likelihood that each datapoint since the hypothetical changepoint belongs to each of the three blocks, and weights by the prior for that block. For simplicity, we assume that the block prior is constant and flat, meaning $P(B) = 1/3$ for all blocks.

The final term, $P(r_{t-1}, \mathbf{x}_{1:t-1})$ is simply the posterior from the previous trial, so we can recursively update the posterior using the estimate from the previous trial, multiplied by the changepoint prior and the predictive probability, appropriately normalized. The probability of a changepoint was defined as the probability density at $r_t = 1$, that is, the probability that a changepoint just occurred.

On each trial, the number of computations grows linearly for each trial, so the model has time complexity $O(N^2)$, which can become costly for long sessions. For this paper, following [20], we implement a modified version that only calculates run-lengths < 75 trials. This modification allows us to run the model with constant time complexity, $O(0)$, and returns essentially equivalent results as the full model since changepoints occur every 40 trials, so we do not expect potential run lengths > 75.

Sterotaxic surgeries

We performed all surgeries using a Neurostar Robot Stererotaxic system on rats after 4 months of age. All rats were induced with 3% isofluorane in oxygen at a flow rate of 2.5 L/minute, which was reduced to 2% isofluorane in oxygen at a flow rat of 1.75 L/minute for maintenance for the duration of the procedure. NAcc injections and implants were targeted to AP 1.3; ML 1.65; DV -6.9 with an 8-10° angle from the midline for bilateral implants.

Photometry

We measured dopamine release using fiber photometry and GRAB_{DA} sensors (AddGene #140554). We injected AAV9-hsyn-GRAB_DA2h to drive expression of the GRAB sensor, as well as AAV1-CB7-CI-mCherry-WPRE-RBG (AddGene #105544) to drive the expression of mCherry to correct for motion artificats. Rats received 60 nL, both delivered over a range of DV values. We implanted 400 μm, 0.5 NA chronically implantable optic fibers (Thorlabs) over

the injection site (DV -6.7 - -6.9). We simultaneously recorded $GRAB_{DA}$ and mCherry fluorescence with Doric Lenses hardware and software (Doric Neuroscience Studio). We corrected for motion artifacts using Two-channel Motion Artifict Correction (TMAC)[55]. For a subset of rats, we corrected for motion artifacts using both the mCherry signal as well as isosbestic recordings of GFP. We found similar results for both methods.

Statistical analyses

Sensitivity to reward blocks

To assess sensitivity to blocks across the population, we z-scored each rat's trial initiation time using the cumulative mean and standard deviation pooled across sessions, and averaged z-scored trial initiation times over blocks. We compared average trial initiation time across blocks using a paired Wilcoxon sign-rank test.

Block transition dynamics

To examine how behaviors changed around block transitions, for each rat, we z-scored their trial initiation times. We removed satiety effects by regressing trial initiation times against trial number and subtracted the fit. We then averaged the z-scored trial initiation times based on their distance from a block transition, including violation trials (e.g., averaged all trials five trials before a block transition). Finally, for each transition type, we smoothed the average transition curve using a causal filter (in order to not introduce pre-transition artifacts) of 10 trials individually for each rat. Finally, we averaged transition curves across rats for each transition type.

Previous reward regression

To capture the trial history effects, we regressed trial initiation time against previous rewards. We focused on mixed blocks only. We linearized the rewards by taking the binary

logarithm of each reward, $\log_2$(reward), and set the reward for unrewarded trials (e.g., violation or catch trials) to 0, since rats do not receive a reward on those trials. We regressed the previous nine $\log_2$(reward) offers, not including the current trial, with a constant offset using Matlab's builtin regress function. With the coefficients, we set the first non-significant coefficient (coefficient whose 90% confidence interval overlapped with 0) and all subsequent coefficients to 0. To quantify the time scale of the coefficients, we fit a negative exponential decay curve, $y = D\exp(-x/\tau)$ to each rat's previous trial coefficients, and reported the time constant (tau) for each rat. If rats had one or fewer significant previous trial coefficients, tau was reported as NaN. For early and late block regressions, we used an identical procedure, but only on the first or last 10 trials of a mixed block.

Photometry

For all photometry analyses, to quantify dopamine release, we measured the AUC of the dopamine response by integrating the dopamine fluorescence from 0 to 0.5 seconds from the event alignment. DA signals were not baseline corrected with the exception of Figure 4M. In that case, baseline was defined as the average response from 0.5 to 0 seconds before delay end, which was subtracted out for each trial. Except where noted, all dopamine analyses were restricted to mixed blocks. To assess reward history effects on dopamine fluorescence, we used similar methods as above, with the inclusion of an additional coefficient for the current trial offer. To compare dopamine AUC to model estimates of reward prediction error for early and late trials (first or last 10 trials), we used the RPE estimates fit to the respective trial type and the dopamine responses only on those trials. Then for each individual session, we regressed dopamine response against RPE separately for positive and negative RPEs, given the rectification of negative RPE encoding, using Matlab's builtin robustfit function.

Acknowledgments

We thank members of the Constantinople lab for feedback and helpful discussions. We thank research technicians in the Constantinople lab for animal training.

Funding: This work was supported by R00MH111926, an Alfred P. Sloan Research Fellowship, a Klingenstein-Simons Fellowship in Neuroscience, and an NIH Director's New Innovator Award (DP2MH126376) to C.M.C. C.G. was supported by a grant from the Simons Foundation (855332)), F32MH125448, and 5T32MH019524. A.M. was supported by 5T90DA043219, F31MH130121, and 5T32MH019524.

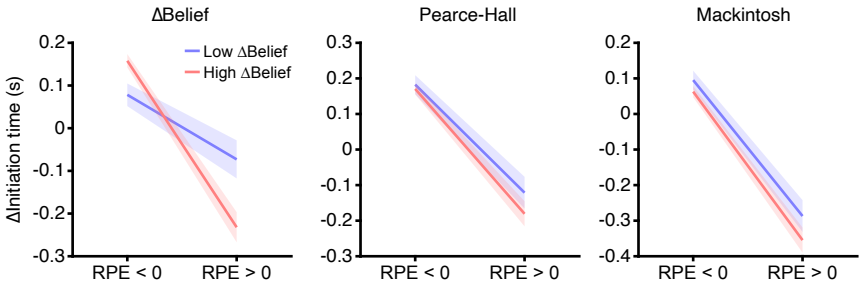
Author Contributions

C.M.C. and A.M designed the study. C.E.G. collected the photometry data and contributed to photometry analysis. A.M. developed the behavioral model and performed behavioral and photometry analyses. A.M. prepared the figures. C.M.C. and A.M. wrote the manuscript. C.M.C. supervised the project.

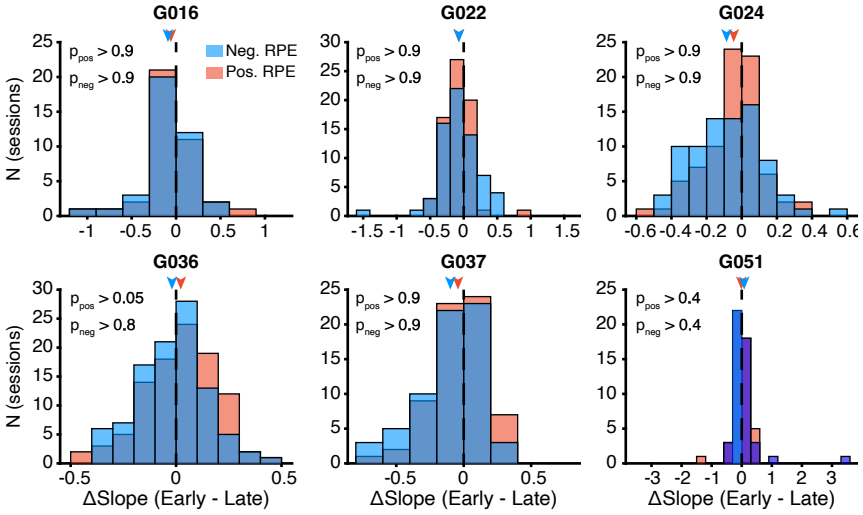
Declaration of Interests

The authors declare no competing interests.

Supplemental Material



Supp. Fig. 1: **Only ΔBelief model captures rat behavior.** Change in initiation time for positive and negative RPE bin conditioned on change in belief for each model. Data are mean ± S.E.M.



Supp. Fig. 2: **RPE vs. DA slopes are comparable early and late in mixed blocks across rats** Change in slope for early vs. late DA vs. RPE regression, fit separately for positive and negative RPEs for each sessions, for individual rats. Arrows indicate mean.