Nucleus accumbens dopamine release reflects Bayesian inference during instrumental learning

Albert J. Qü[1,2 *], Lung-Hao Tai[3], Christopher D. Hall[4], Emilie M. Tu[1], Maria K. Eckstein[5], Karyna Mishchanchuk[6], Wan Chen Lin[3], Juliana B. Chase[1], Andrew F. MacAskill[6], Anne G.E. Collins[1,3], Samuel J. Gershman[7,8], and Linda Wilbrecht[1,3]

[1]Department of Psychology, University of California, Berkeley, CA, 94720, USA
[2]Center for Computational Biology, University of California, Berkeley, CA, 94720, USA
[3]Helen Wills Neuroscience Institute, University of California, Berkeley, CA, 94720, USA
[4]Sainsbury Wellcome Centre for Neural Circuits and Behaviour, University College London, London, W1T 4JG, UK
[5]Google DeepMind, London, UK
[6]Department of Neuroscience, Physiology and Pharmacology, University College London, UK
[7]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, MA, 02138, USA
[8]Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[*]Correspondence: albert_qu@berkeley.edu

Abstract

Dopamine release in the nucleus accumbens has been hypothesized to signal reward prediction error, the difference between observed and predicted reward, suggesting a biological implementation for reinforcement learning. Rigorous tests of this hypothesis require assumptions about how the brain maps sensory signals to reward predictions, yet this mapping is still poorly understood. In particular, the mapping is non-trivial when sensory signals provide ambiguous information about the hidden state of the environment. Previous work using classical conditioning tasks has suggested that reward predictions are generated conditional on probabilistic beliefs about the hidden state, such that dopamine implicitly reflects these beliefs. Here we test this hypothesis in the context of an instrumental task (a two-armed bandit), where the hidden state switches repeatedly. We measured choice behavior and recorded dLight signals reflecting dopamine release in the nucleus accumbens core. Model comparison based on the behavioral data favored models that used Bayesian updating of probabilistic beliefs. These same models also quantitatively matched the dopamine measurements better than non-Bayesian alternatives. We conclude that probabilistic belief computation plays a fundamental role in instrumental performance and associated mesolimbic dopamine signaling.

Introduction

Reinforcement learning (RL) algorithms hypothesize that, in value based decision making tasks, animals maintain an action value map and update it using reward prediction errors (RPE), the difference between the observed and predicted reward associated with the chosen action [1]. The growing family of RL models has proved remarkably successful for explaining trial-by-trial changes in behavior [2–4] and the responses of dopamine neurons in rodents and primates [5–7]. However, these standard RL (SRL) models may not account as well for behavioral and neural data when sensory observations are not sufficient to determine optimal behavior, for example, when the underlying reward distribution changes over time due to the existence of a hidden state. To solve such tasks, evidence suggests that the brain represents a "belief state" (a probability distribution over hidden states), updated by Bayesian inference [8–12], or possibly learned implicitly in an end-to-end fashion [13].

The significance of Bayesian inference becomes apparent when studying mouse behavior in a two-armed bandit task (2ABT, Fig. 1A), a serial reversal learning task where rewards for correct choices are delivered on a probabilistic schedule, and the correct port switches between two available options across blocks of trials within a single session. The fact that rewards are delivered on a probabilistic schedule generates ambiguity as to which port is currently the rewarded port following an unrewarded trial. Both mice and human subjects trained in this task are capable of maintaining stable behavior within a block, and then rapidly adapting after reward contingencies reverse following a block switch [14–16]. The simplest versions of standard RL models do not capture choice stability within a block and

rapid switching when the rewarded port changes [15, 16].

Bayesian inference may offer better models of behavior in tasks like the 2ABT because they capture stable within-block choice and rapid between-block switching [11]. However, the additional explanatory power of Bayesian models with respect to RL models is currently ambiguous. Comparison of Bayesian and RL models using human behavior in the 2ABT suggested that they both provide complementary explanations of behavior in the 2ABT (provided the RL models are relatively sophisticated, as explained below), but neither was decisively superior [15]. The limitations of a purely Bayesian account of reversal learning was also challenged by recent behavioral analyses of mice performing the 2ABT [16]. Here we report data from mice performing the 2ABT along with recording of dopamine release events in the nucleus accumbens core (NAc) using fiber photometry imaging of dLight [17] . These data allow us to compare Bayesian and SRL models to examine which computations mice use to solve the 2ABT.

In addition to comparing "pure" Bayesian and SRL models, it is also valuable to include a hybrid model, as past studies suggest that Bayesian and RL processes may be combined in the brain [8, 18, 19]. Therefore, we also examined a belief state RL model (BRL), where Bayesian inference is used to compute belief states, over which RL processes operate to learn policies appropriate for the current belief state. Past studies have shown that dopamine signaling may be better explained by models accounting for belief states [8–10, 12, 20]. We hypothesized that RPE signals generated by the BRL model would match dopamine measurements more closely than RPE signals generated by alternative models. We also hypothesized that the BRL model would better match the choice data, in particular the rapid adaptation to block switches exhibited by mice.

Results

Mice rapidly adapt to block switches in the 2ABT

To study flexible updating of goal-directed behavior under probabilistic conditions, we trained mice on a 2ABT task with block switches. Two weeks prior to training, mice were injected bilaterally with AAV dLight in the NAc core and implanted with an optic fiber and ferrule above the injection site.

In the 2ABT, mice encountered three ports equipped with infrared sensors. Animals were water restricted and trained to nose-poke into the central port to initiate a trial and then move to a left or a right port to obtain a water reward. Water rewards were available on either the left or right side, depending on the block. The correct choice was rewarded with water 75% of the time whereas no rewards were available at the other port. After they received a random number of rewards (uniformly sampled from 7 to 23), the correct port switched without any discriminative cue. In order to achieve a desirable amount of water rewards across the whole session, mice needed to constantly update their choice strategies

through outcome feedback.

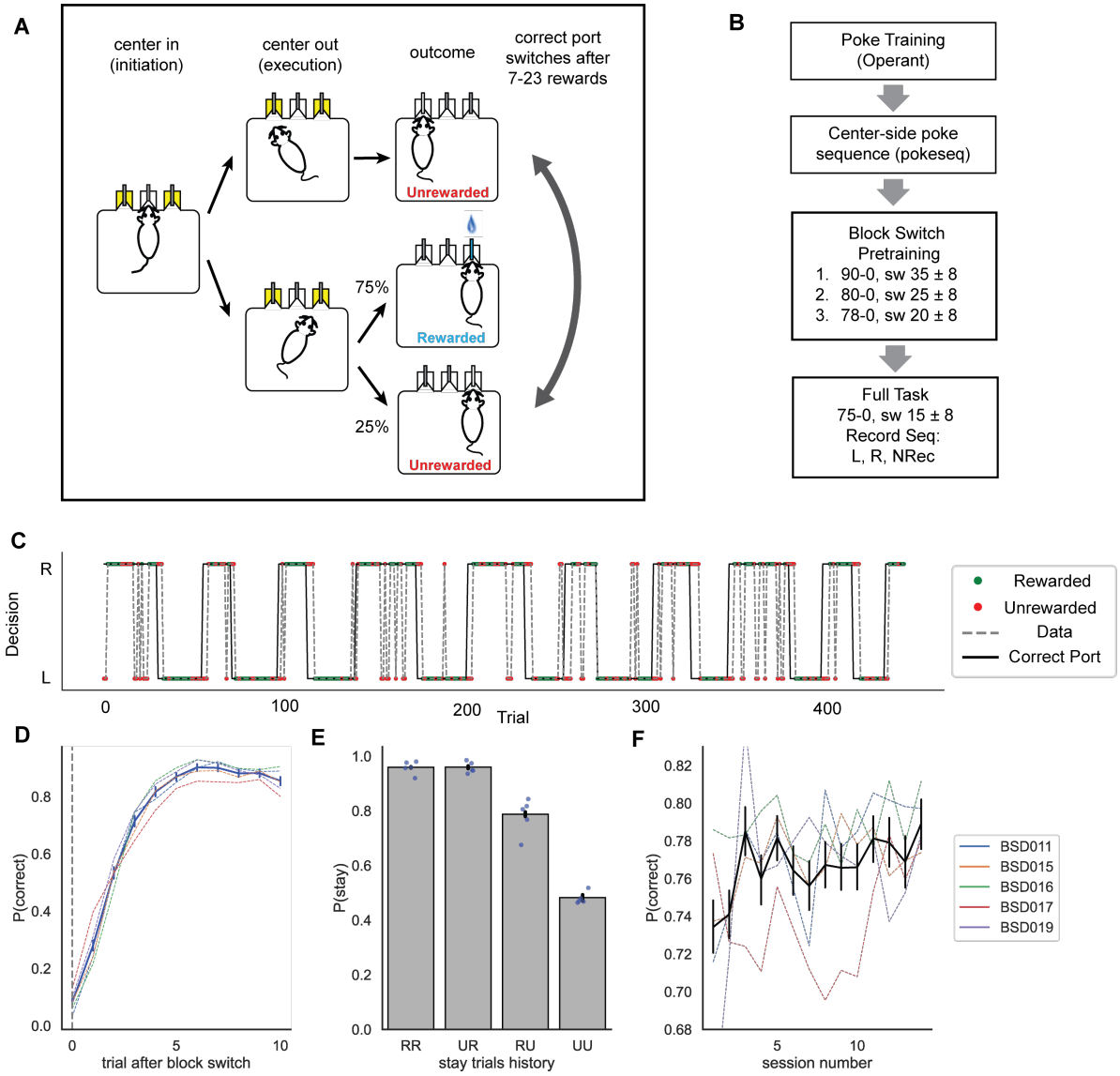
After pre-training (see Methods for details), we recorded unilaterally from the NAc in mice performing the 2ABT for a total of 14 daily sessions, alternating the hemisphere and allowing one non-recording session every third day (Fig. 1B). Just after pre-training, mice chose the high reward probability port 72.9% (95% CI: 0.669, 0.767) of the time on average. Over the course of 14 training sessions, performance steadily increased (Fig. 1F, linear regression slope coefficient 0.0026, t statistic: 2.743, p=0.008, CI: 0.001, 0.005). In the first 7 sessions, mice took 2.49 trials on average (95% CI: 2.39, 2.59) to switch to the correct port after a block switch. In the 8-14th session this reduced to an average of 2.36 trials (95% CI: 2.28, 2.45). Mice switched and committed to the correct port for 3 or more trials after an average of 3.12 trials (95% CI: 3.00, 3.26) following the first 7 sessions. These data were comparable to mice performing the task without a fiber implant or cable [14].

To identify what task features the mice might consider before switching ports, we analyzed the stay/switch behaviors after two trials of outcome history. To simplify the comparison, we specifically chose the trials after the mice consecutively stayed in the same port for two trials or more. We encoded the outcome history at t-2, t-1 as RR, RU, UR, RU, (23607, 5719, 9428, 8520 trials each) with R representing rewarded trials and U representing unrewarded trials. For instance, a trial with "RU" stay history means that a mouse encountered a reward at trial t-2 and then got unrewarded at trial t-1, all at the same port. We found a significant reduction in the probability of repeating the previous choice after one unrewarded outcome (RR vs. RU, Fisher's exact test: 6.47, P<1e-4, Cohen's d: 0.6242, 95% CI: 0.6, 0.65). Moreover, mice reduced their stay rate at the chosen port after two consecutive unrewarded outcomes compared to one unrewarded outcome (RU vs. UU, Fisher's exact test: 3.98, P<1e-4, Cohen's d: 0.6715, 95% CI: 0.64, 0.7; Fig 1E).

Cognitive model fitting confirms that Bayesian models recover mouse behavior better than RL models

To investigate possible computations that mice use to rapidly adapt their choices following block switches, we compared a Bayesian model (BIfp, see methods) adopted from [15], a BRL model (BRLfwr, see methods) inspired by [8], and several SRL models (Fig. 2): a model with asymmetric learning rates for positive and negative RPEs (RL4p), a model that additionally used counterfactual updates for unchosen option values (RLCF), and the recursively formulated logistic regression (RFLR) model developed by [16]. The RFLR model was grouped with other SRL models due to its mathematical equivalence to the forgetting Q-learning model [16, 21]. Modeling details can be found in the Methods and descriptions shown in Table 1. A summary of the model relationships is shown in Fig. 2A.

All the compared models compute choice values, map these to choice probabilities, and then update their values (and beliefs in the case of the Bayesian models) after receiving reward feedback. Only the RL models (both SRL and BRL) use RPEs to update values, but we

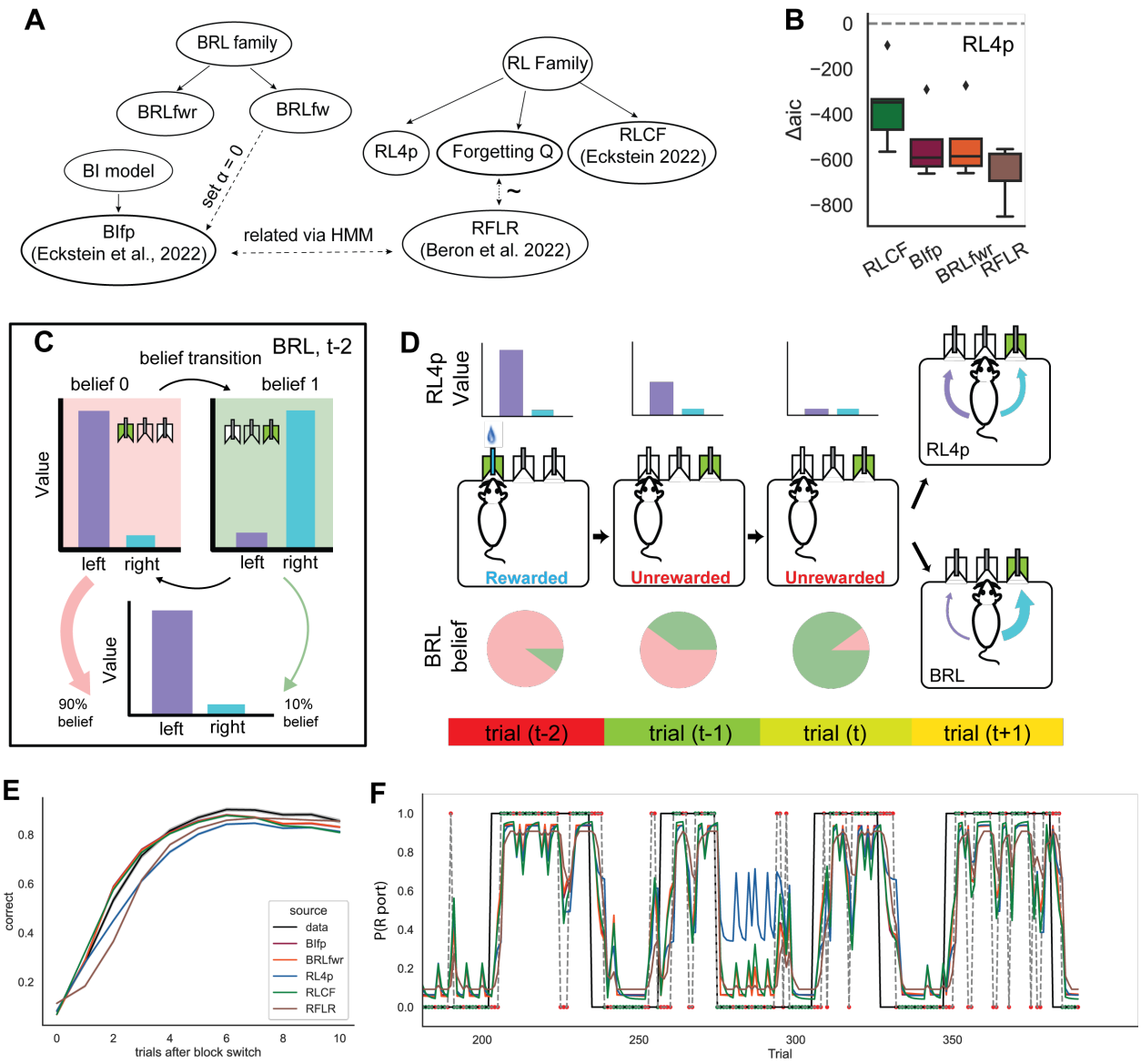

A
center in
(initiation)
center out
(execution)
outcome
correct port
switches after
7-23 rewards
Unrewarded
75%
Rewarded
25%
Unrewarded
B
Poke Training
(Operant)
Center-side poke
sequence (pokeseq)
Block Switch
Pretraining
1. 90-0, sw 35 ± 8
2. 80-0, sw 25 ± 8
3. 78-0, sw 20 ± 8
Full Task
75-0, sw 15 ± 8
Record Seq:
L, R, NRec
C
Decision
R
L
0
100
200
Trial
300
400
Rewarded
Unrewarded
Data
Correct Port
D
P(correct)
trial after block switch
E
P(stay)
RR
UR
RU
UU
stay trials history
F
P(correct)
session number
BSD011
BSD015
BSD016
BSD017
BSD019

model name	description	parameters
RL4p	Q learning	α^+, α^-, β, ϕ (stickiness)
RLCF	Q learning with counterfactual updates [15]	α^+, α^-, β, ϕ
RFLR	Recursive Formulation Logistic Regression [16]	α, ϕ, τ (decay rate)
BRLfwr	Belief state RL model with reward weight updates and fixed initial weight	β, ϕ, q (latent switch rate), α
BRLfw	Belief state RL model with full weight updates	β, ϕ, q, α
BIfp	Bayesian Inference Model [22]	β, ϕ, q

Table 1: Overview of different cognitive models

can calculate "pseudo-RPEs" for the remaining models by taking the difference between the observed and expected reward. Critically, the BRL and SRL models differ in how they perform value computation. The SRL models we considered update choice values stored in a look-up table. The BRL models computed values as weighted linear functions of the belief

Figure 1 *(preceding page)*: **Mice adapt rapidly to block switches in a probabilistic reversal task.** (A) Illustration of the two-armed bandit task, divided into initiation, execution, and outcome phases. In the illustrated trial, the right port is rewarded with 0.75 probability and the left port is unrewarded. After 7-23 rewarded trials, the correct port switches. (B) Training protocol. The recording phase took place in the "Full Task" phase. In the pretraining phases, the structure of the task was the same as the in the full task phase, except the reward contingencies and block lengths were different. Each contingency is labeled by numbers indicating the proportion of correct and incorrect choices that were rewarded. For example, "90-0" in the first pretraining phase indicates that 90% of correct choices were rewarded. The block length in each phase is indicated by its mean and range. For example, "sw 35 ± 8" in the first pretraining phase indicates that switches occurred after the animal earned between 27 and 43 rewards. During the 14 sessions of mice behavior data collection, we recorded dLight signals using a "left hemisphere (L), right hemisphere (R), no neural recording pure behavior (NRec)" sequence. (C) Raw behavioral trajectory taken from the first half of a sample session. Black line indicates correct reward port locations while dashed gray line indicates actual mouse behavior. Green dots and red dots mark rewarded and unrewarded trials, respectively. (D) Probability of making a correct choice (i.e., choosing the high probability port) as a function of the number of trials after a block switch. Each dashed line shows trajectories for individual animals. (E) Probability of staying (repeating the last choice) after experiencing different outcome histories in the same port. RR: two consecutive rewards; UR: unrewarded outcome followed by rewarded; RU: rewarded outcome followed by unrewarded; UU: two consecutive unrewarded outcomes. (F) Performance across 14 sessions. Dashed lines show individual animal trajectories. Error bars show 95% bootstrapped confidence intervals.


A
BRL family
BRLfwr
BRLfw
BI model
set α = 0
BIfp
(Eckstein et al., 2022)
related via HMM
RL Family
RL4p
Forgetting Q
RLCF
(Eckstein 2022)
~
RFLR
(Beron et al. 2022)
B
RL4p
Δaic
0
−200
−400
−600
−800
RLCF
BIfp
BRLfwr
RFLR
C
BRL, t-2
belief transition
belief 0
belief 1
Value
left
right
90%
belief
10%
belief
D
RL4p
Value
Rewarded
Unrewarded
Unrewarded
RL4p
BRL
BRL
belief
trial (t-2)
trial (t-1)
trial (t)
trial (t+1)
E
correct
trials after block switch
source
data
BIfp
BRLfwr
RL4p
RLCF
RFLR
F
P(R port)
Trial

state, where the weights can be thought of as standard RL choice values conditioned on the belief states. The updating process is schematized in Fig. 2C-D.

The RL4p model and its variants have been extensively validated in previous studies (e.g., [23–27]), which have provided evidence for both asymmetric learning rates and choice perseveration ("stickiness"). This model serves as a baseline against which all more complex models should be compared. Despite its past empirical success, a critical limitation of the RL4p model is that it fails to capture the observation that animals appear to update values for unchosen options (counterfactual updating; [15, 28–30]). For example, in reversal learning tasks like the 2ABT, observing an unexpected reward omission dramatically increases the likelihood of a switch, accompanied by neural responses that anticipates the new reward contingencies [20, 31]. Counterfactual updating is usually formalized by updating values for unchosen actions in the opposite direction from the values of chosen actions. This qualitatively mimics the behavior of Bayesian models, which also naturally give rise to counterfactual updating.

We fitted all models to the choice data of all 14 sessions for 5 mice in the 2ABT using maximum likelihood estimation. To perform model selection, we computed the Akaike Information Criterion (AIC) using the log-likelihood values for each model. With RL4p as a baseline, we found that both BRLfwr and BIfp significantly reduced the total AIC measure compared to the RL4p model (BIfp: $\Delta\text{AIC} = -536.57 \pm 59.48$, BRLfwr: $\Delta\text{AIC} = -530.59 \pm 61.86$). The RLCF also explained the mouse behavior better than the RL4p model (RLCF: $\Delta\text{AIC} = -361.92 \pm 70.52$), with a higher relative AIC on average than the BIfp model ($\Delta\text{AIC} = 174.64 \pm 47.83$) (Fig 2B). Additionally, RFLR model also outperformed the RL4p model ($\Delta\text{AIC} = -652.53 \pm 49.26$). We concluded that either BIfp, BRLfwr, or

Figure 2 *(preceding page)*: **Bayesian and reinforcement learning models.** (A) Relationships between models. We focused on a particular form of Bayesian (BI) model, denoted BIfp (see Methods for more details). This can be derived as a variation of the Bayesian reinforcement learning (BRL) models by setting the learning rate to 0. The standard reinforcement learning (SRL) family contains the RL4p, RLCF and Forgetting Q models (the latter is approximately equivalent to the RFLR model used in the paper). RFLR is related to the BIfp model as it approximates inference in a hidden Markov model (HMM) under restricted conditions [16]. Interestingly, through the use of counterfactuals, RLCF also approximate a "pseudo-belief" with its Q values. (B) Model comparison using relative AIC compared to RL4p. (C) Illustration of value computation for BRL model family, which updates beliefs via Bayes' rule and then uses these beliefs to compute values. (D) Illustration using a four-trial sequence to show the differences between RL4p and BRL. Top: purple and cyan bars show the choice values; Bottom: pie charts show the belief state for BRL. (E) Behavior of different models compared to mouse data (black line). Trial 0 is when the computer has switched the rewarded side in a block switch. (F) Example behavioral trajectory (probability of choosing the rightward port) predicted by different models. Mouse data are marked by a dashed line and block structure is marked by a solid line. Rewarded trials are marked as green dots and unrewarded trials are marked as red dots.

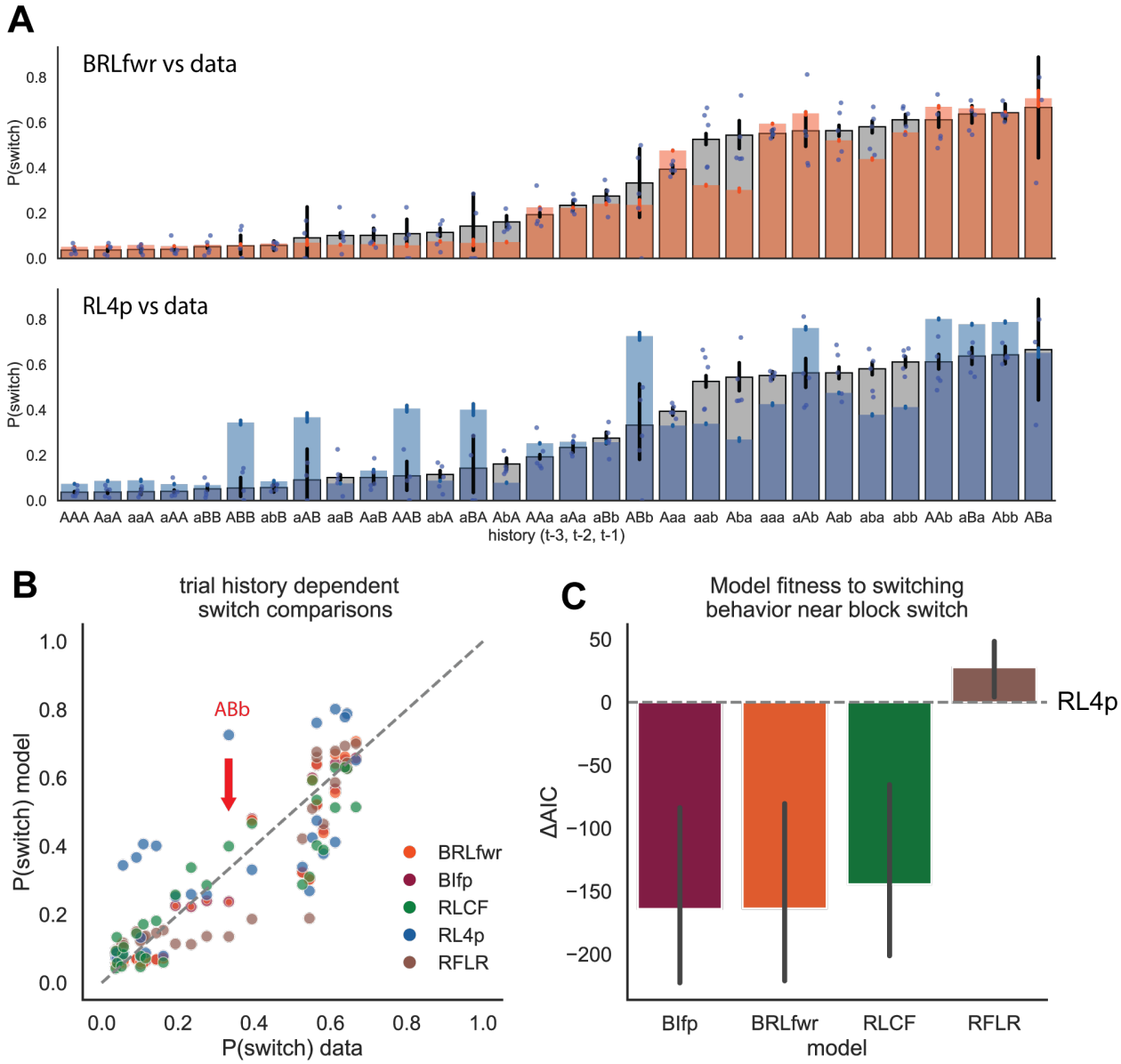
more complex SRL models like RLCF and RFLR, provided better accounts of mouse behavior in the 2ABT on all trials overall. There was not sufficient discriminative power in the data to identify which of these models was superior, although the RLCF model fared slightly worse than the others.

Bayesian models qualitatively recover mouse behaviors following task block switches

To further understand the mechanism by which the Bayesian models explain mouse behavior, we compared qualitative signatures of the BIfp, BRLfwr, and SRL models with those of mouse data. One hallmark behavioral signature of mouse adaptation to block switches is improved choice accuracy as the number of trials after a block switch increases (Fig. 1D). Accordingly, we first compared the rate at which choice accuracy improved after block switches for BIfp, BRLfwr, RL4p, RLCF and RFLR against mouse data. Consistent with our hypothesis, we observed that BRLfwr and BIfp predict faster adaptation to block switches compared to RL4p, resembling the adaptation rate of mouse switching behavior. Interestingly, BIfp and BRLfwr were qualitatively indistinguishable (Fig. 2E). When we focus on single trial choice-outcome trajectories, we qualitatively observed that the two Bayesian models updated their switch probability differently from the SRL models after unrewarded observations or near block switches. From this, one further hypothesis naturally arises: the BIfp and BRLfwr models outperform SRL models because of their ability to detect choice-outcome sequences that are suggestive of a block switch via Bayesian inference.

To test this hypothesis, it is imperative to obtain a more granular understanding of how different trial outcome histories cause a mouse to switch ports, and to determine if BIfp and BRLfwr can successfully predict switch probabilities under different outcome histories, especially around block switches. Following [16], we categorized all trials based on their three past trial outcome histories, using capital A/B to denote rewarded outcomes and lower-case a/b for unrewarded outcomes. Due to the symmetrical task structure, we denoted the trial at t-3 as A/a regardless of its spatial location, and trials at t-2 or t-1 as A/a if the mouse chose the same port, or B/b if the mouse chose a different port, in reference to trial t-3 (Fig. 3A). For instance, if a mouse was unrewarded at port 1 at t-3, rewarded at port 2 at t-2, unrewarded again at port 2 at t-1, we would describe the trial outcome history as aBb. We then calculated the switch probability as the probability to choose a different port at trial t compared to trial t-1.

We then compared the average mouse switch probability after each outcome history against the mean predictions of the three different cognitive models (Fig. 3B; BIfp Sum of Squared Errors, SSE: 0.1724, BRLfwr: 0.1796, RL4p SSE: 0.8270, RLCF SSE: 0.2748, RFLR: 0.3352). Notably, we observed that the RL4p model overestimated the switch probability for outcome history contexts where mice encountered rewards in both ports within the past 3 trials (e.g. ABb, AaB, etc.). Since these outcome contexts occur only around block switches, mice usually stayed at the newly rewarded ports (mean switch rate: 0.2143). Bayesian models


A
BRLfwr vs data
RL4p vs data
P(switch)
0.0
0.2
0.4
0.6
0.8
AAA AaA aaA aAA aBB ABB abB aAB aaB AaB AAB abA aBA AbA AAa aAa aBb ABb Aaa aab Aba aaa aAb Aab aba abb AAb aBa Abb ABa
history (t-3, t-2, t-1)
B
trial history dependent
switch comparisons
ABb
P(switch) model
P(switch) data
BRLfwr
BIfp
RLCF
RL4p
RFLR
C
Model fitness to switching
behavior near block switch
ΔAIC
50
0
−50
−100
−150
−200
RL4p
BIfp
BRLfwr
RLCF
RFLR
model

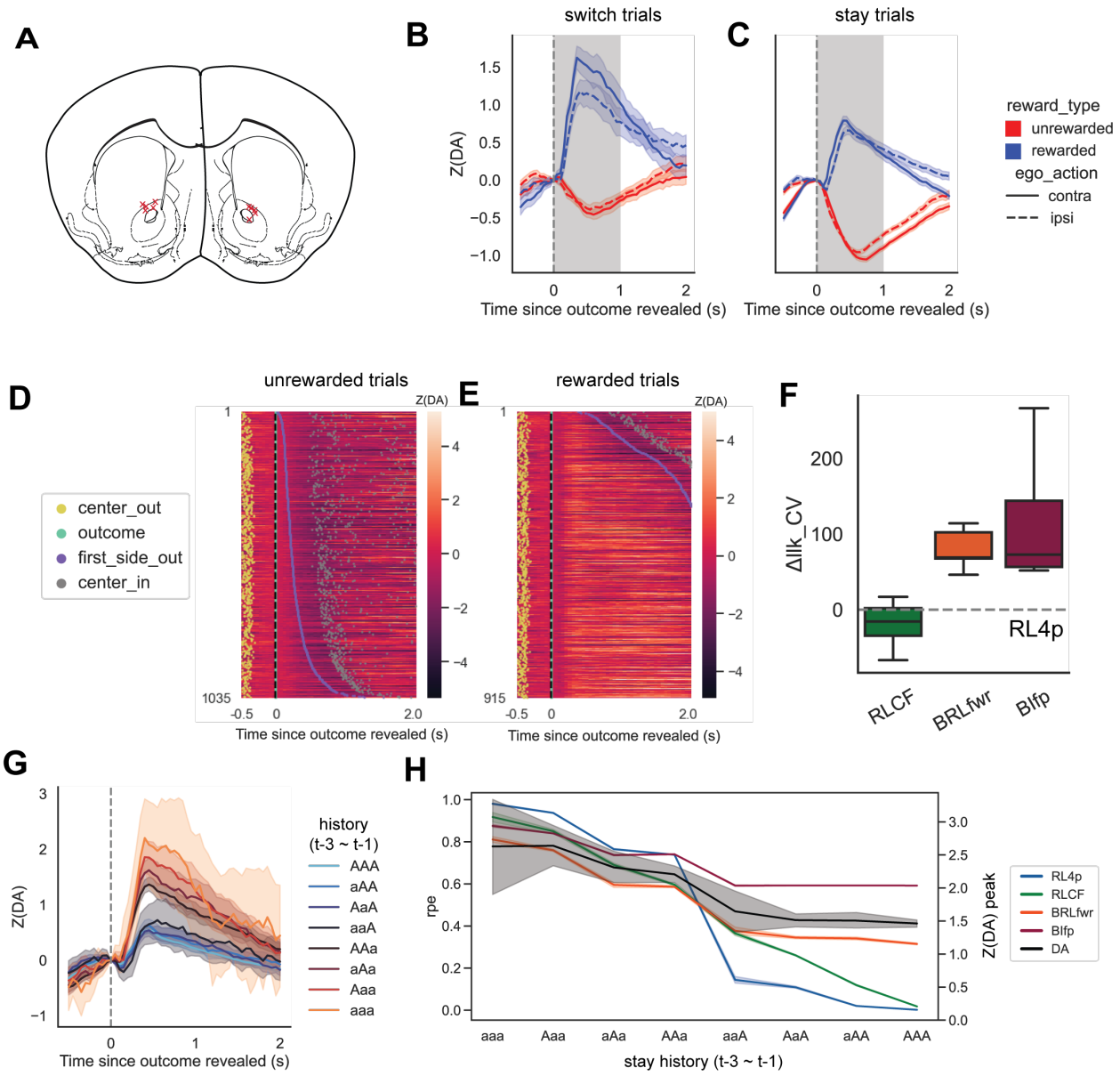
showed similarly low switch rates after these contexts (BIfp: 0.1740, BRLfwr: 0.1801), whereas RL4p model predicted a high average switch rate of 0.4328 (RLCF: 0.2248, RFLR: 0.2159). For instance, after the outcome context of ABb, the RL4p model overestimated the probability of the mice switching back to port A (Fig. 3B), because it was only able to increment the action value for port B, instead of capturing the underlying reward distribution change. A mouse using a learning mechanism like RL4p would mistakenly think that selecting port A was still as valuable of an action as it was prior to this outcome sequence given that no reward omissions were recently experienced at that port, leading to regressive errors back to port A.

To investigate this phenomenon further, we selected the first 5 trials after block switches for each block (including the trial at which block switches). Then we compared the relative AIC for both the BI, BRL and the SRL models, with RL4p again as a baseline (Fig. 3C). We found that BIfp and BRLfwr indeed explained mouse behavior better than RL4p (BRLfwr: ΔAIC $= -163.68 \pm 35.92$, BIfp: ΔAIC $= -163.99 \pm 36.47$). Interestingly, RLCF (ΔAIC $= -144.32 \pm 34.65$) also explained mouse behavior well in trials near block switches. With counterfactual updates, Q values of the two alternative actions always sum up to one, approximating the belief distribution similar to Bayesian models. However, due to the differences in update mechanisms, we hypothesize that there are still differences in RPE predictions between RLCF and Bayesian models that dopamine recording data can further dissociate.

Predictions from BRLfwr and BIfp capture Nucleus Accumbens core dLight dynamics better than SRL models

We next investigated if either BIfp, BRLfwr or SRL models provide accurate quantitatively accurate predictions of dopamine release in the nucleus accumbens triggered by rewards. Importantly, BIfp does not inherently calculate an RPE term since it simply uses Bayesian inference for model updates. However, we noted that mathematically BIfp operates equivalently to a BRL model with its model weights fixed throughout model updates. Therefore, to allow the comparison between model predictions of BIfp and the dopamine signals, we leveraged this equivalence to calculate a pseudo-RPE for BIfp.

Figure 3 *(preceding page)*: **BRL outperforms standard RL by better explaining mouse behaviors around block switches.** (A) Switch probability by different trial outcome histories described by action-outcome pairs three trials back. Gray bars showed mouse average probability of switching for each outcome history, deep blue dots represent individual mice. Top: Mouse data overlayed with BRLfwr model predictions of switch rate in orange; bottom: mouse data overlayed with RL4p model predictions of switch rate in blue. (B) Switch probability predicted by different models scales with probability of mice switching port selections in different outcome contexts described. Colors represent different models, sharing the same legend as C (orange: BRLfwr, wine red: BIfp, dark green: RLCF, brown: RFLR, blue: RL4p) C. Relative AIC with respect to RL4p (dashed line at ΔAIC $= 0$) show model fit to mouse data around block switch. Error bars show 95% bootstrapped confidence intervals.


A
B
switch trials
C
stay trials
Z(DA)
Time since outcome revealed (s)
reward_type
unrewarded
rewarded
ego_action
contra
ipsi
D
unrewarded trials
E
rewarded trials
center_out
outcome
first_side_out
center_in
F
ΔIlk_CV
RL4p
RLCF
BRLfwr
BIfp
G
history
(t-3 ~ t-1)
AAA
aAA
AaA
aaA
AAa
aAa
Aaa
aaa
H
rpe
Z(DA) peak
stay history (t-3 ~ t-1)
DA

As mentioned above, our experimental mice had been implanted with optical fibers bilaterally in NAc core to enable recording of dLight signals. Histology images were visually inspected after the experiment to verify the implant tip location and viral expression (Fig. 4A). We analyzed the dopamine responses in NAc by aligning dopamine signals to the "outcome" event, the time point when water rewards were either delivered to the mice or were omitted. After mice received a water reward, we observed a dLight dopamine signal increase on average in ports contralateral and ipsilateral to the recording hemisphere, consistent with previous literature [6]. When reward was omitted at the peripheral port, we observed a reduction in the dLight signal in NAc (Fig. 4B-C).

We also observed a significant difference in trial-averaged dLight responses to the "outcome" event between switch trials and stay trials. Signals triggered by reward were larger on average on trials where mice switched to a new port, compared to when they stayed with their past choice. Negative signals observed after reward omissions in trials where mice stayed were larger when compared to when they had just switched (Fig. 4B-C).

When we visualized single trial dopamine traces around "outcome" events, we found that the temporal span of the dopamine response closely aligned with the amount of time mice stay at the peripheral reward port (Fig. 4E). Similarly, after unrewarded outcomes, the decreased dLight response span tightly corresponded to the time mice stayed in the peripheral reward port, reaching signal trough typically just after mice left the port. These results together suggest that, when modeling dopamine responses, the time duration that mice stayed in the peripheral reward port after an "outcome" event (dubbed "port duration") need to be taken

---

Figure 4 *(preceding page)*: **NAc dLight dopamine dynamics consistent with belief state RL model RPE predictions.** (A) Implant fiber locations indicated on mouse brain atlas with red crosses. (B-C) Trial average of NAc dLight signals (z-scored, as described in Methods) aligned to outcome events. Shaded area indicates the one second where the peak or trough is taken for neural regression. B shows switch trials and C shows stay trials. Rewarded trials are in blue and unrewarded trials are in red. Trials where mice picked the port contralateral to the recording hemisphere are plotted with solid lines; trials in which mice picked the ipsilateral port are plotted with dashed lines. (D-E) Example session single trial dLight responses plotted in heatmaps, trials sorted by the time mice spent in the reward port (see Methods for further details). D shows a heatmap for unrewarded trials, and E shows rewarded trials. Increase in dLight signal is indicated by brighter shades of red and decreases from baseline are indicated by darker shades of black. Dots are used to mark "center out" (yellow), "outcome" (green), "first side out" (purple), "center in" (gray) events, respectively. (F) Result of neural regression using model RPE values to explain dopamine variability. Fit is measured as cross validated log-likelihood (llk_CV) relative to the RL4p model. Gray dashed line indicates the baseline of RL4p RPE fitted to dopamine measurements. (G) Dopamine response on rewarded trials binned by past history, sorted in increasing order of number and recency of rewards (note in all cases mice stayed with the same port 'a/A' for all three trials). (H) RPE predictions from different models plotted against dopamine peak values (in black). Error bars show 95% bootstrapped confidence intervals.

into consideration. Additionally, after controlling for staying or switching decisions, there were subtle differences in the trial-average dLight responses depending on the laterality of the chosen port relative to hemisphere recorded (dubbed "egocentric action"; Fig. 4D-E).

If learning mechanisms of the mice are computed in the NAc and resemble that of BRLfwr or BIfp, the RPE calculation given by them should explain a higher amount of variance in the dopamine responses in NAc at the time of "outcome" events compared to the SRL models. To simplify the testing procedure, we obtained dopamine summary statistics by taking the peak of dopamine transients for rewarded trials and troughs for unrewarded trials (dubbed DA-PT) during the one second window after the outcome, marked in gray in Fig. 4B. To control for confounding variables, we included session number, port duration, egocentric action, movement time, and center port poke duration as covariates in addition to predicted RPE values from BIfp, BRLfwr, RL4p, and RLCF (note that we omitted RFLR model since it does not have a straightforward RPE formulation), each in a disjoint regression covariate set. Using maximum likelihood estimation with featurized covariates (see Methods) on the five different RPE covariate sets, we found that both BIfp and BRLfwr models yielded larger relative log-likelihood (Δllk CV) than the RL4p baseline, evaluated on cross validation sets (70-30) for each animal (Fig. 4F, BRLfwr Δllk CV: 79.87, 95% CI: [59.60, 102.51], BIfp Δllk CV: 118.34, 95% CI: [65.39, 217.50]). Both models also outperformed RLCF on this measure (RLCF Δllk CV: -19.62, 95% CI: [-47.40, 3.64]).

To further understand what neural phenomenon the BRL and BIfp models captured that the SRL models failed to capture, we looked at the changes in dLight responses to rewards as mice experienced a series of rewards in consecutive stay trials from aaa to AAA (lowercase "a" indicates unrewarded and uppercase "A" indicates rewarded). As mice encountered more rewards in their recent history, dLight dopamine responses to reward gradually decreased (consistent with decreased RPE) but did not completely disappear (Fig. 4G).

This pattern observed in our dLight data from the NAc core was not well captured by the SRL models (Fig. 4H). These models update action values on each trial as mice encounter consecutive rewards, estimating a recent average of choice outcomes. This feature may make them overly sensitive to recent outcome histories and liable to overestimate action values, particularly around block switches. In addition, RL4p models require large α values (learning rate) to maintain the flexibility to adapt to block switches. In our dataset, the best fitting RL4p model had an average α^+ of 0.85 (95% CI: [0.77, 0.94]), and average α^- of 0.73 (95% CI: 0.71, 0.75). As a result, after two or three consecutive rewards, mice using an RL4p mechanism would associate the selected port with an action value close to 1, resulting in an RPE value close to zero, which is not what we observed.

Both BIfp and BRLfwr were able to explain this phenomenon in the dLight signals better than SRL models. The BIfp and BRLfp considered possible block switches and maintained values of both a high reward context and a low reward context that are not associated with any given port. Instead, these values were assigned to specific ports in proportion to the model's belief state on any given trial. This information about task structure provided

robustness to probabilistic reward omissions after selecting the more rewarding port. This also allowed the best fitting BRLfp and BIfp model to predict a small but non-negative RPE response on rewarded stay trials as there is always a degree of uncertainty as to whether the block had switched (Fig. 4H), capturing a pattern observed in the NAc dLight signals.

Discussion

Our results showed that mice were capable of rapidly adapting to block switches in an instrumental task with probabilistic and volatile contingencies, without sacrificing the robustness of their choice policy when encountering stochastic unrewarded outcomes after correct choices. Both the Bayesian inference model and the belief state RL model using Bayesian inference were able to capture this balance and outperformed standard RL models in explaining mouse behavior and dopamine release events in the NAc core following choice outcomes. Critically, both Bayesian models had explicit belief state representations of different block identities, which provided a mechanism to encode the rapidly changing reward contingencies in the 2ABT. The model successfully explained the ability to efficiently switch after consecutive unrewarded outcomes while also disregarding occasional reward omissions within a block. The RL component in the BRL models allows the model to use RPEs to update its mapping from choices to values conditioned on belief states. However, we did not find decisive evidence that BRL models outperform pure Bayesian models (which use a fixed mapping) in terms of matching mouse behavior. The principal conceptual advantage of BRL models is that they make direct use of RPE signals, whereas pure Bayesian models do not. This highlights the importance of neural data in adjudicating between these models.

Building on the behavioral modeling, we showed that the same models do a good job explaining NAc dopamine activity at the time of outcomes (reward or reward omission). In particular, RPEs generated by the BRL model closely match the history-dependent pattern of dopamine activity. This finding supports a growing literature showing that dopamine signaling of RPEs depends on a belief state (or some approximation of a belief state; see [13]) in partially observable tasks [8–12].

The BRL model is closely related to structure learning models that learn multiple context-dependent policies. Such models often assume some form of Bayesian inference at the more abstract level, with RL supporting learning of specific policies [18, 32, 33]. However, theoretical work, supported by experimental findings, have also shown that dynamic adaptation at the more abstract level of belief states/latent contexts/rules may also be supported by RL-like computations [22, 34, 35]. The current work cannot dissociate those possibilities.

We believe several limitations of this present work should be addressed in future investigations. First, the 75-0 reward probability design implies that, in any given block, a low probability port always yields no reward. Consequently, mice rarely switch to a new port and get rewarded after observing rewards in the previously chosen port. Previous work has noted that inference-based cognitive models are capable of decreasing the expectations about

alternative options after observing a reward in selected option [20, 31]. Therefore, mice using inference would exhibit a higher RPE following a reward in a new port after also observing a reward in the old port, denoted as R-SW-R (reward-switch-reward) sequence. Yet in our dataset, there are only 94 R-SW-R trials out of 30707 trials in total, across multiple animal sessions (0.3%), giving us limited power to address this prediction. In addition, in our task design, mice can initiate new trials at their discretion after the outcome, with no instructed delays. Especially after unrewarded outcomes, mice leave the peripheral port shortly after, approaching the center port for a new trial. As a result, in trials when mice leave the peripheral port too soon, dopamine responses coincided temporally with the dopamine ramp associated with approaching center ports [36], which can also be explained as an aggregating sequence of RPE signal as animals reach towards a goal (in this case, the initiation of next trial) [37, 38]. To mitigate this, we only analyzed trials with specific port durations (see methods), which limited our statistical power. This also poses the general challenge of teasing out the contributions of the two temporal components.

Another limitation of our study is that the RPE that correlates with DA signaling only plays a minimal role in predicting behavior, as evidenced by the fact that BRL models did not outperform pure Bayesian models with no RL component. We believe that better discriminating these possibilities will require theory-guided experimental design, possibly using a more complex task that places a greater demand on learning policies within each belief state [39, 40].

Conclusion

In natural environments, animals and humans often experience ambiguous outcome feedback about the hidden reward structure of the environment. We emulated this in a two-armed bandit task with switching reward blocks, and showed that mice are capable of adapting rapidly to changes in hidden reward structures after observing specific outcome feedback sequences. Cognitive model fitting to behavioral data suggested that models performing Bayesian updating of beliefs better explained mouse behavioral data than various standard RL models. These Bayesian models also quantitatively and qualitatively matched the dopamine release in nucleus accumbens core better than the non-Bayesian alternatives. Together, we conclude that probabilistic belief updates are critical to behavioral adaptation and RPE signaling in the mesolimbic dopaminergic system during instrumental learning in a partially observable environment.

Acknowledgements: We are grateful to Scott Linderman, Bernardo Sabatini, Celia Beron, Jing Jing Li, Thomas W. Elston, Hongli Wang, Gaia Molinaro, Jaclyn Essig, and Eric J. Hu for discussion and to Chris Machle, Lexi Z. Zhou and Anthony T. Dunn for discussion and assistance with data collection. This work was supported by the National Institutes of Health (grant number U19NS113201 to L.W. and S.J.G.) and the Air Force Office of Scientific Research (grant number FA9550-20-1-0413 to S.J.G.).

Author Contributions: Conceptualization, M.K.E., A.G.E.C., S.J.G., C.D.H., A.J.Q., K.M., A.F.M., L.W. ; Methodology, S.J.G., A.J.Q., C.D.H., A.G.E.C, M.K.E, L.T., L.W., W.C.L. ; Software, A.J.Q., S.J.G. ; Validation, K.M., A.F.M., C.D.H., L.T. ; Formal Analysis, A.J.Q., S.J.G. ; Investigation, A.J.Q., C.D.H., E.M.T., L.T. ; Resources, L.T., L.W. ; Data Curation, A.J.Q., E.M.T., L.T., J.B.C. ; Writing - Original Draft, A.J.Q., E.M.T., L.W., S.J.G., C.D.H. ; Writing – Review & Editing, A.G.E.C., S.J.G., L.W., C.D.H., M.K.E., A.J.Q. ; Visualization, A.J.Q., S.J.G., A.G.E.C., L.W., C.D.H., K.M., A.F.M., M.K.E., W.C.L. ; Supervision, L.W., S.J.G., A.G.E.C. ; Project Administration, A.J.Q., C.D.H., L.W., and S.J.G. ; Funding Acquisition, L.W., S.J.G.

Methods

Animal protocol

Mice (all C57 Bl/6 male, bred in-house aged 104-167 days) were housed on a 12 h reversed light-dark cycle (lights on at 22:00) with nesting material. Prior to surgery they were group housed with access to food and water ad libitum. We conducted all animal procedures, which follow the principles outlined by the NIH Guide for the Care and Use of Laboratory Animals, according to the protocols approved by the University of California, Berkeley Institutional Animal Care and Use Committee (IACUC) and Office of Laboratory Animal Care (OLAC).

Probabilistic Switching 2ABT Task

We trained 5 male mice on a 2ABT behavioral task in which the location of the water reward was periodically switched between the two potential rewarded ports at random intervals. Trials were initiated by a nose poke in the center port and concluded after the mouse subsequently chose one of two reward ports, located on either side of the center port. Nose pokes were detected by a infrared photodiode. Once the initiating poke was sensed, lights that encompass the reward ports were triggered to cue the animals to the viability of a potential water reward. The mouse receives a 2 μl water reward or no reward depending on their correct or incorrect, respectively, port choice for each trial.

Only one peripheral port was rewarded at a time. The setting for the number and frequency of water rewards was altered during each phase of the task. Teaching animals the task, we started them in the "operant" phase, where water rewards were offered in the peripheral ports, and then moved to the "pokeseq" phase, where mice learned to poke the middle port before going to retrieve a peripheral port water reward. Successful learning during this phase was measured as the mouse getting more than 700 rewards in less than five hours. Some food pellets were placed into the arena while the mice were learning the task to motivate them. We omitted mice from the experiment if they didn't perform successfully on "pokeseq" after five days.

Once the mice learned to poke the center port to initiate the peripheral port water rewards, we no longer put food into the arenas and advance the mice onto the "learning switch" phase. During the three subphases of the learning switch, we teach the mice that only one port is rewarded at a frequency of 90%, 80%, and 78% for each subphase respectively. Additionally, we taught the mice that the rewarded port switches after obtaining a random number of cumulative rewards set within the ranges of 27-43, 17-33, and 12-28 for each subphase respectively. Water rewards became less predictable and more variable with each subsequent phase.

Successful learning for the three learning switch phases were set at a score higher than 69.5% on both the left and right port within four, three, and two hours for each subphase respectively. If the mice seemed unmotivated, we placed some food into the arena. Once

they completed the task with food, we removed the food and had them reach criterion for their appropriate subphase.

Fibers were first introduced to the mice after they complete the last learning switch subphase and are utilized for neural recording only after the mice successfully reach the learning switch criteria once again. During the "full task", used for fiber photometry recording, mice received a water reward with a 75% chance if they picked the correct port, but received nothing if they picked the wrong port for each trial. Once mice obtained a random number of cumulative rewards, sampled from 7 to 23 uniformly, from one port, the reward was switched to the other port. This structure minimized the possibility for the mice to predict the timing of the switch.

When participating in this task, the mice were water restricted but maintained ad libitum access to food. They were supplemented with additional water if they earned less than 500 µl water from the task. Mice were weighed daily, before and after the task, and were given additional water if needed, at least 30 minutes after training sessions, to ensure that they maintained around 85% of their original weight.

Surgery protocol

Mice were anesthetized with isoflurane gas for stereotaxic surgery. Meloxicam was given on the day of surgery and daily for 48h after. Coordinates for the Nucleus Accumbens Core were bregma coordinate: 1.20mm anterior, ± 1.2mm medial-lateral, -4.1mm ventral. AAV-CAG-dLight1.3b or AAV9-syn-dLight1.2b were injected using a Nanojet II (Drummond scientific). Neurophotometrics NA 0.37 or 0.48 400μm optic fibers were placed approximately 100μm above the injection site. Dental cement was used to secure the implant to the skull. During a recovery period of 7-14 days, mice were singly-housed and fed ad libitum.

Histology and Imaging

Following behavior, animals were transcardially perfused with 4% paraformaldehyde (PFA) in an 0.1M phosphate buffer (PB) solution (pH = 7.4). Brains were collected and fixed overnight, followed by a transfer to 0.1M PB. To visualize striatal photometry fibers, perfused brains were sliced coronally at 50μm using a vibratome (VT1000S Leica Biosystems; Buffalo Grove, IL). Immunohistochemistry was performed to amplify the GFP signal (1:1000 chicken anti-GFP, Aves Labs, Inc.; GFP-1020 followed by 1:1000 goat anti-chicken AlexaFluor 488, Invitrogen by Thermo Fisher Scientific; A11039). Slides were mounted on slides with Fluoromount-G (Southern Biotech). Slices anterior and posterior to the fiber tract were imaged at 10X on an AxioScan Z.1 fluorescent microscope (CRL Molecular Imaging Center, UC Berkeley) to confirm targeting.

Task Behavior Analysis

All final behavioral analysis is conducted with custom implemented python packages. Behavioral data are first collected using a custom implemented 2ABT control module in matlab, and then preprocessed into behavioral data frames with each row representing a separate trial with various columns containing information regarding task meta data, mice choice data, and task relevant variables. For trials that mice failed to initiate and make a choice, we included the data but noted the relevant variables as NaNs/null. For most behavioral analysis, outcome histories as well as covariates with null entries are dropped. During cognitive modeling, for null trials, we simplied maintained the model latent and skipped the model update steps. All model plotting are generated via `seaborn` packages, and test statistics and various measures are generated using `statsmodels, pingouin and sklearn`.

We divided the tasks into initiation phase, execution phase, outcome phase and inter trial interval (ITI) phase. The initiation phase includes both a "center in" and "center out" event when mice poke their nose into and out of the center port to initiate a trial. To execute a choice, mice leave the center port and then enter one of two peripheral ports to make a selection. When the outcome of their choice is revealed to the mice, they linger at the port for water reward or time-out until they leave the port during the "side out event".

We defined the following critical events:

- Center in (CI): when mice first poke into the center port to start a new trial.
- Center out (CO): when mice leave the center port to select ports.
- Side in (SI): when mice poke peripheral port.
- Outcome (O): when outcome is revealed, typically immediately after side in (latency on the order of system latency).
- Zeroth side out (SO0): first time when mice move the nose poke out of the port trigger IR beam break.
- First side out (SO1): last time of a consecutive set of beam breaks at the same side peripheral port. Zeroth side out and first side out can often be the same when the animal leaves the port only after drinking all the water.
- Last side out (SOf): last side out beam breaks before next center port poke to initiate a new trial.

We defined the following critical timing measures:

- Movement time (`MVMT`) : CI - SOf (t-1)
- Inter-trial interval (`ITI`): CI - SO1(t-1)

- Center duration (`center_dur`): CO-CI
- Port duration (`port_dur`): SO1-O
- Side out bout latency (`SO_lat`): SO1-SO0, typically close to 0, critically this is used to control for atypical trials with excessive beam breaks.

Computational Modeling

We formalize the decision problem facing animals as follows. At time t, the animal makes a choice $c_t \in \{-1, 1\}$ and then observes a reward $r_t \in \{0, 1\}$. By convention, we take $c_t = -1$ to denote the left port, and $c_t = 1$ to denote the right port. The reward probability depends on the chosen action and the hidden state $z_t \in \{-1, 1\}$:

$$P(r_t = 1 | c_t, z_t) = \begin{cases} \rho_1 & \text{if } c_t = z_t \\ \rho_2 & \text{if } c_t \neq z_t \end{cases} \tag{1}$$

where ρ_1 is the probability of reward for a correct choice, which equals 0.75 during the phase when photometry measurements were collected, and ρ_2 is the probability of reward for an incorrect choice, which is technically 0 but which we set to 0.0001 to allow for model misspecification. The hidden state changes on each trial with probability $q = P(z_t \neq z_{t-1})$.

We assume a common functional form for the choice policy across models:

$$P(c_t = 1) = \sigma(\beta D_t + \phi c_{t-1}) \tag{2}$$

where $\sigma(x) = 1/(1 + e^{-x})$, $D_t = Q_t(1) - Q_t(-1)$ is the value difference, $\beta \geq 0$ is an inverse temperature parameter, $\phi \geq 0$ is a stickiness (choice perseveration) parameter, and $Q_t(c)$ is the value assigned to choice c at time t. The stickiness component captures the tendency to repeat choices independent of the reward history.

The models make different assumptions about how the values are updated based on experience:

- **RL4p** is a standard RL model that updates Q-values according to a delta rule, with separate learning rates for positive (α^+) and negative (α^-) reward prediction error (RPE):

$$\Delta Q_t(c_t) = \alpha_t \delta_t \tag{3}$$

$$\alpha_t = \begin{cases} \alpha^+ & \text{if } \delta_t \geq 0 \\ \alpha^- & \text{if } \delta_t < 0 \end{cases} \tag{4}$$

  where $\delta_t = r_t - Q_t(c_t)$ is the RPE.

- **RLCF** is an elaboration of the RL4p model that uses "counterfactual" updates [15]. Whereas RL4p only updates the chosen action value, RLCF additionally updates the

unchosen action value in the opposite direction:

$$\Delta Q_t(-c_t) = \alpha_t \tilde{\delta}_t \tag{5}$$

where $-c_t$ is the unchosen action and $\tilde{\delta}_t = 1 - r_t - Q_t(-c_t)$ is the counterfactual RPE.

- **BRL** (belief state reinforcement learning) is an RL algorithm that computes the values as a linear function of the belief state $b_t(z) = P(z_t = z | c_{1:t-1}, r_{1:t-1})$, the posterior probability over the hidden state given the choice and reward history (the belief state update will be described further below):

$$Q_t(c) = w_1 b_t(c_t) + w_2 b_t(-c_t), \tag{6}$$

where w_1 and w_2 are learnable weights. Intuitively, the expected reward is decomposed into two terms: the first term captures the reward probability when the animal is correct ($z_t = c_t$), and the second term captures the reward probability when the animal is incorrect ($z_t = -c_t$). If the animal has perfectly learned the task, the weights should be dictated by the ground truth reward probabilities, $w_1 = \rho_1$ and $w_2 = \rho_2$. We formalize a model of learning (**BRLfw**) based on gradient descent, which does not require the animal to know the ground truth:

$$\Delta w_1 = \alpha \delta_t b_t(c_t) \tag{7}$$

$$\Delta w_2 = \alpha \delta_t b_t(-c_t), \tag{8}$$

where α is a learning rate. We considered two versions of the model. The weights were initialized to $w_1 = \rho_1$ and $w_2 = \rho_2$ (note that choosing other initial conditions had relatively little effect on model fits). Note that the weights can be interpreted as representing the Q-values of correct actions conditioned on the belief, with their update rule a confidence-weighted RL update (see e.g. [41]), hence the name BRL. We also considered a restricted version of the model (**BRLfwr**) where we fix $w_2 = \rho_2$, which we found to perform just as well as the unrestricted version, so we focused on the restricted version in the main results. Belief state updating followed Bayes' rule, with $P(z|\tilde{z})$ denoting the transition probability of the hidden state from $\tilde{z}$ to z:

$$b_{t+1}(z) \propto \sum_{\tilde{z} \in \{-1,1\}} P(z|\tilde{z}) b_t(\tilde{z}) P(r_t | c_t, \tilde{z}) \tag{9}$$

We assume that the belief state is initialized to $b_1(z) = 0.5$. Also we parameterize $q = P(z \neq \tilde{z} | \tilde{z})$ as the transition probability parameter of the model.

- **BIfp** (Bayesian inference with fixed parameters) makes use of the same belief state as the BRL models, but where there is no updating of the weights. Instead, these are fixed to $w_1 = \rho_1$ and $w_2 = \rho_2$. Interestingly, the best fitted parameter for α in the unrestricted BRL version (BRLfw) above was zero for most subjects, establishing an equivalence, under this 2ABT task, between BIfp and and BRLfw.

- **RFLR** (recursively formulated logistic regression; [16]) is a form of RL that is mathematically equivalent to a form of logistic regression. RFLR updates the value difference variable according to:

$$D_{t+1} = e^{-\frac{1}{\tau}} D_t + \alpha r_t c_t, \tag{10}$$

  where τ is a timescale parameter governing the exponential decay rate of the values. Notably, we removed the non-negativity constraint on ϕ and fixed $\beta = 1$ in equation 2 for RFLR specifically, as it is consistent with the definition in [16] and yields slightly better performance.

Model fitting and comparison: We fit all models to 14 sessions of all 6 mice during the Probswitch phase using custom implemented `cogmodels` python module. The module finds the best-fitting model parameters for each animal by maximum likelihood estimation. The model initializes the parameters by randomly sampling from distributions as follows:

- RL4p: $\alpha^+ \in \text{Unif}(0,1),\ \alpha^- \sim \text{Unif}(0,1),\ \phi \sim \Gamma(2, 0.2),\ \beta \sim \text{Exp}(1)$
- RLCF: Same as RL4p
- RFLR: $\alpha \sim \text{Exp}(1),\ \phi \sim \mathcal{N}(0,1),\ \tau \sim \text{Exp}(1)$
- BIfp: $\beta \sim \text{Exp}(1),\ \phi \sim \Gamma(2, 0.2),\ q \sim \text{Unif}(0, 0.05)$
- BRLfwr: Same parameters as BIfp, plus $\alpha \sim \text{Unif}(0,1)$

where $\Gamma(\alpha, \theta)$ denotes the gamma distribution with shape α and scale β; $\mathcal{N}(\mu, \nu^2)$ denotes the Gaussian distribution with mean μ and variance ν^2.

After obtaining maximum likelihood estimates, we constructed an empirical range of parameters and performed model recovery tests. We simulated behavioral data from randomly sample model parameters for each model from the empirical ranges, and refit to the simulated data to obtain the model-fitted values for 1000 random iterations. All three model classes are able to recover model parameters with high correlation (Figure S2). We performed model comparison by computing the Akaike Information Criterion (AIC) for each model $\mathcal{M}_i$:

$$\text{AIC} = 2k - 2\sum_t \big(\ln P(c_t|\hat{\theta}, \mathcal{M}_i)\big), \tag{11}$$

where $P(c_t|\hat{\theta}, \mathcal{M}_i)$ is the likelihood of the data c_t conditional on the maximum likelihood estimate ($\hat{\theta}$). We report AIC scores relative to RL4p, the baseline RL model. More negative scores indicate that a model explains the data better compared to RL4p.

Fiber Photometry (FP)

We recorded from all 5 mice with fiber implants in nucleus accumbens core (NAc) using Neurophotometrics FP3001 system at 20Hz. dLight signals are measured in 470nm channel,

while isosbestic control signal is measured in 415nm channel for artifact removal. During the learning switch phase, we start plugging in fiber implant without recording to get mice accustomed to moving and behaving with fibers. Starting from the first "full task" session, we alternate between a left hemisphere recording, right hemisphere recording, and no recording schedule. All FP recordings are synchronized to behavioral and video data via custom TTL systems and bonsai programs and saved for further processing. Recorded signals are preprocessed using custom python program to control for motion artifacts using 415nm reference channels [42]. Specifically, we first subtract the common DC shift by approximating a smooth trend across both channels. Then we perform a robust linear regression from reference channel to signal channel, and subtract the fitted 415nm artifact signal from 470nm channel recordings and obtain the processed dLight signals. All dLight signals are subsequently z-scored to account for across-session extraneous variations.

Neural Data Analysis

All analysis was carried out with a custom Python module for aligning, visualization, and data modeling. After baseline correction, we visually inspected both 415nm channel recording and 470nm channel recording, as well as the AUC-ROC scores between the rescaled baseline values against signal channel recordings. We picked sessions that meets the following conditions: AUC-ROC higher than 0.9, longer tail in distributions in 470nm recording compared to rescaled 415nm recordings suggesting clear fluorescence increases, and no sharp discontinuity in both channels' recordings. We use baseline-corrected and z-scored dLight signals for in-depth analysis, called Z(DA). For neural data alignment, we interpolated Z(DA) around event times and appended them to the trial level dataframe.

Aligned neural signals are appended to the trial level data frames as additional columns. Aligned Z(DA) is baselined by subtracting out its value at "outcome" time (we denote this procedure as de-base for simplicity). For trial average dopamine visualization, we first identified a list of relevant columns as grouping variables of interest, we then dropped any row with one or more NaN entries in the relevant columns. The trial averaged plots are then plotted using seaborn with the error bar being bootstrapped confidence intervals. When reporting summary statistics, we used bootstrapped confidence intervals to characterize mean estimates. For difference comparisons, we used suitable paired or unpaired tests and reported both the testing statistics as well as effect size, and the confidence intervals associated with it.

For neural data regression, we first calculated a summary measure of dopamine in the outcome phase. With Z(DA) aligned to "outcome" and de-based, we calculated the peak of rewarded trial responses, and trough of unrewarded trial responses. To control for different time scales of port stays after outcome, we filtered out trials with latency between `first_side_out` and `zeroth_side_out` higher than 0.5s. The covariates we included for the regression analysis are session number, egocentric action, MVMT, center duration, as well as model RPE, and port duration. We did a logarithmic transform for movement time

and ITI, and dropped trials with movement time longer than 3s, `port_dur` longer than 6s, or `center_dur` longer than 0.8s. In addition, we dropped trials where `SO1 - SO0 > 1s`. Then we encoded RPE and port durations for rewarded trials and unrewarded trials separately. All features are normalized using the z-score and verified to have VIF under 3 before regression. To compare results for different model RPEs, we obtained the RPE predictions from best fitted models to behavioral data for all trials, and appended the RPE as an additional feature to the trial behavioral structures, along with behavioral variables like `port_dur` or dopamine summary statistics. We divided the datasets into training and testing sets in 70:30 ratio. For each cognitive model, we trained an OLS regressor taking in aforementioned covariates, and calculated the loglikelihood of observing the `DA_PT` values in the held-out testing set, denoted as `llk_cv`. Here we reported the relative `llk_cv` with respect to RL4p.

References

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction.* Second edition. Adaptive computation and machine learning series. Cambridge, Massachusetts: The MIT Press, 2018. ISBN: 978-0-262-03924-6.

[2] Jan Gläscher, Alan N. Hampton, and John P. O'Doherty. "Determining a Role for Ventromedial Prefrontal Cortex in Encoding Action-Based Value Signals During Reward-Related Decision Making". en. In: *Cerebral Cortex* 19.2 (Feb. 2009), pp. 483–495. ISSN: 1460-2199, 1047-3211. DOI: 10.1093/cercor/bhn098. URL: https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bhn098 (visited on 09/27/2023).

[3] Tobias U. Hauser et al. "Cognitive flexibility in adolescence: Neural and behavioral mechanisms of reward prediction error processing in adaptive decision making during development". en. In: *NeuroImage* 104 (Jan. 2015), pp. 347–354. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2014.09.018. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811914007605 (visited on 09/27/2023).

[4] Samantha R. Santacruz et al. "Caudate Microstimulation Increases Value of Specific Choices". en. In: *Current Biology* 27.21 (Nov. 2017), 3375–3383.e3. ISSN: 09609822. DOI: 10.1016/j.cub.2017.09.051. URL: https://linkinghub.elsevier.com/retrieve/pii/S0960982217312496 (visited on 09/27/2023).

[5] Pr Montague, P Dayan, and Tj Sejnowski. "A framework for mesencephalic dopamine systems based on predictive Hebbian learning". en. In: *J. Neurosci.* 16.5 (Mar. 1996), pp. 1936–1947. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.16-05-01936.1996. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.16-05-01936.1996 (visited on 07/30/2023).

[6] Wolfram Schultz, Peter Dayan, and P. Read Montague. "A Neural Substrate of Prediction and Reward". en. In: *Science* 275.5306 (Mar. 1997), pp. 1593–1599. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.275.5306.1593. URL: https://www.science.org/doi/10.1126/science.275.5306.1593 (visited on 04/19/2023).

[7] Neir Eshel et al. "Dopamine neurons share common response function for reward prediction error". en. In: *Nat Neurosci* 19.3 (Mar. 2016), pp. 479–486. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.4239. URL: https://www.nature.com/articles/nn.4239 (visited on 09/23/2023).

[8] Benedicte M. Babayan, Naoshige Uchida, and Samuel. J. Gershman. "Belief state representation in the dopamine system". en. In: *Nat Commun* 9.1 (May 2018), p. 1891. ISSN: 2041-1723. DOI: 10.1038/s41467-018-04397-0. URL: https://www.nature.com/articles/s41467-018-04397-0 (visited on 04/17/2023).

[9] Clara Kwon Starkweather et al. "Dopamine reward prediction errors reflect hidden-state inference across time". en. In: *Nat Neurosci* 20.4 (Apr. 2017), pp. 581–589. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.4520. URL: https://www.nature.com/articles/nn.4520 (visited on 09/23/2023).

[10] Samuel J. Gershman and Naoshige Uchida. "Believing in dopamine". en. In: *Nat Rev Neurosci* 20.11 (Nov. 2019), pp. 703–714. ISSN: 1471-003X, 1471-0048. DOI: 10.1038/

s41583-019-0220-7. URL: http://www.nature.com/articles/s41583-019-0220-7 (visited on 04/17/2023).

[11] Vincent D. Costa et al. "Reversal Learning and Dopamine: A Bayesian Perspective". en. In: *J. Neurosci.* 35.6 (Feb. 2015), pp. 2407–2416. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.1989-14.2015. URL: https://www.jneurosci.org/lookup/doi/10.1523/JNEUROSCI.1989-14.2015 (visited on 08/08/2023).

[12] Armin Lak et al. "Midbrain Dopamine Neurons Signal Belief in Choice Accuracy during a Perceptual Decision". en. In: *Current Biology* 27.6 (Mar. 2017), pp. 821–832. ISSN: 09609822. DOI: 10.1016/j.cub.2017.02.026. URL: https://linkinghub.elsevier.com/retrieve/pii/S096098221730163X (visited on 09/23/2023).

[13] Jay A Hennig et al. "Emergence of belief-like representations through reinforcement learning". In: *PLOS Computational Biology* 19.9 (2023), e1011067.

[14] Lung-Hao Tai et al. "Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value". en. In: *Nat Neurosci* 15.9 (Sept. 2012), pp. 1281–1289. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.3188. URL: http://www.nature.com/articles/nn.3188 (visited on 04/18/2023).

[15] Maria K. Eckstein et al. "Reinforcement learning and Bayesian inference provide complementary models for the unique advantage of adolescents in stochastic reversal". en. In: *Developmental Cognitive Neuroscience* 55 (June 2022), p. 101106. ISSN: 18789293. DOI: 10.1016/j.dcn.2022.101106. URL: https://linkinghub.elsevier.com/retrieve/pii/S1878929322000494 (visited on 04/17/2023).

[16] Celia C. Beron et al. "Mice exhibit stochastic and efficient action switching during probabilistic decision making". en. In: *Proc. Natl. Acad. Sci. U.S.A.* 119.15 (Apr. 2022), e2113961119. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2113961119. URL: https://pnas.org/doi/full/10.1073/pnas.2113961119 (visited on 04/21/2023).

[17] Tommaso Patriarchi et al. "Ultrafast neuronal imaging of dopamine dynamics with designed genetically encoded sensors". In: *Science* 360.6396 (2018), eaat4422. DOI: 10.1126/science.aat4422. eprint: https://www.science.org/doi/pdf/10.1126/science.aat4422. URL: https://www.science.org/doi/abs/10.1126/science.aat4422.

[18] Anne Collins and Etienne Koechlin. "Reasoning, Learning, and Creativity: Frontal Lobe Function and Human Decision-Making". In: *PLOS Biology* 10.3 (Mar. 2012), pp. 1–16. DOI: 10.1371/journal.pbio.1001293. URL: https://doi.org/10.1371/journal.pbio.1001293.

[19] Milena Rmus, Samuel D McDougle, and Anne Ge Collins. "The role of executive function in shaping reinforcement learning". en. In: *Current Opinion in Behavioral Sciences* 38 (Apr. 2021), pp. 66–73. ISSN: 23521546. DOI: 10.1016/j.cobeha.2020.10.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620301480 (visited on 09/26/2023).

[20] Ethan S. Bromberg-Martin et al. "A Pallidus-Habenula-Dopamine Pathway Signals Inferred Stimulus Values". en. In: *Journal of Neurophysiology* 104.2 (Aug. 2010), pp. 1068–

1076. ISSN: 0022-3077, 1522-1598. DOI: 10.1152/jn.00158.2010. URL: https://www.physiology.org/doi/10.1152/jn.00158.2010 (visited on 06/08/2023).

[21] Makoto Ito and Kenji Doya. "Validation of decision-making models and analysis of decision variables in the rat basal ganglia". In: *Journal of Neuroscience* 29.31 (2009), pp. 9861–9874.

[22] Maria K. Eckstein and Anne G. E. Collins. "Computational evidence for hierarchically structured reinforcement learning in humans". en. In: *Proc. Natl. Acad. Sci. U.S.A.* 117.47 (Nov. 2020), pp. 29381–29389. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1912330117. URL: https://pnas.org/doi/full/10.1073/pnas.1912330117 (visited on 09/26/2023).

[23] Eldad Yechiam et al. "Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits". In: *Psychological science* 16.12 (2005), pp. 973–978.

[24] Michael J Frank et al. "Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning". In: *Proceedings of the National Academy of Sciences* 104.41 (2007), pp. 16311–16316.

[25] Yael Niv et al. "Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain". In: *Journal of Neuroscience* 32.2 (2012), pp. 551–562.

[26] Samuel J Gershman. "Do learning rates adapt to the distribution of rewards?" In: *Psychonomic Bulletin & Review* 22 (2015), pp. 1320–1327.

[27] Michiyo Sugawara and Kentaro Katahira. "Dissociation between asymmetric value updating and perseverance in human reinforcement learning". In: *Scientific Reports* 11.1 (2021), p. 3574.

[28] Erie D Boorman, Timothy E Behrens, and Matthew F Rushworth. "Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex". In: *PLoS Biology* 9.6 (2011), e1001093.

[29] Stefano Palminteri et al. "The computational development of reinforcement learning during adolescence". In: *PLoS Computational Biology* 12.6 (2016), e1004953.

[30] Stefano Palminteri et al. "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing". In: *PLoS computational biology* 13.8 (2017), e1005684.

[31] Alan N Hampton, Peter Bossaerts, and John P O'doherty. "The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans". In: *Journal of Neuroscience* 26.32 (2006), pp. 8360–8367.

[32] Maël Donoso, Anne G. E. Collins, and Etienne Koechlin. "Foundations of human reasoning in the prefrontal cortex". en. In: *Science* 344.6191 (June 2014), pp. 1481–1486. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1252254. URL: https://www.science.org/doi/10.1126/science.1252254 (visited on 09/26/2023).

[33] James B. Heald, Máté Lengyel, and Daniel M. Wolpert. "Contextual inference underlies the learning of sensorimotor repertoires". en. In: *Nature* 600.7889 (Dec. 2021), pp. 489–493. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-021-04129-3. URL: https://www.nature.com/articles/s41586-021-04129-3 (visited on 09/26/2023).

[34] Anne G. E. Collins and Michael J. Frank. "Cognitive control over learning: Creating, clustering, and generalizing task-set structure." en. In: *Psychological Review* 120.1 (Jan. 2013), pp. 190–229. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/a0030852. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030852 (visited on 09/26/2023).

[35] Anne G. E. Collins and Michael J. Frank. "Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory". en. In: *Proc. Natl. Acad. Sci. U.S.A.* 115.10 (Mar. 2018), pp. 2502–2507. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720963115. URL: https://pnas.org/doi/full/10.1073/pnas.1720963115 (visited on 09/26/2023).

[36] Mark W. Howe et al. "Prolonged dopamine signalling in striatum signals proximity and value of distant rewards". en. In: *Nature* 500.7464 (Aug. 2013), pp. 575–579. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12475. URL: http://www.nature.com/articles/nature12475 (visited on 04/22/2023).

[37] HyungGoo R. Kim et al. "A Unified Framework for Dopamine Signals across Timescales". en. In: *Cell* 183.6 (Dec. 2020), 1600–1616.e25. ISSN: 00928674. DOI: 10.1016/j.cell.2020.11.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0092867420315300 (visited on 09/23/2023).

[38] John G Mikhael et al. "The role of state uncertainty in the dynamics of dopamine". In: *Current Biology* 32.5 (2022), pp. 1077–1087.

[39] Thomas Akam, Rui Costa, and Peter Dayan. "Simple Plans or Sophisticated Habits? State, Transition and Learning Interactions in the Two-Step Task". In: *PLOS Computational Biology* 11.12 (Dec. 2015), pp. 1–25. DOI: 10.1371/journal.pcbi.1004648. URL: https://doi.org/10.1371/journal.pcbi.1004648.

[40] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. "Dorsal hippocampus contributes to model-based planning". en. In: *Nat Neurosci* 20.9 (Sept. 2017), pp. 1269–1276. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn.4613. URL: https://www.nature.com/articles/nn.4613 (visited on 11/08/2023).

[41] Kenji Doya et al. "Multiple model-based reinforcement learning". In: *Neural computation* 14.6 (2002), pp. 1347–1369.

[42] Ekaterina Martianova, Sage Aronson, and Christophe D. Proulx. "Multi-Fiber Photometry to Record Neural Activity in Freely-Moving Animals". In: *JoVE* 152 (Oct. 2019), e60278. ISSN: 1940-087X. DOI: 10.3791/60278. URL: https://doi.org/10.3791/60278.