Efficient and stochastic mouse action switching during probabilistic decision making

Celia Beron[1], Shay Neufeld[1], Scott Linderman[2], Bernardo Sabatini[1]

1. Howard Hughes Medical Institute, Department of Neurobiology, Harvard Medical School, Boston, MA
2. Department of Statistics and Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA

Contact info:
bsabatini@hms.harvard.edu
scott.linderman@stanford.edu

Abstract

To gain insight into the process by which animals choose between actions, we trained mice in a two-armed bandit task with time-varying reward probabilities. Whereas past work has modeled the selection of the higher rewarding port in such tasks, we sought to also model the trial-to-trial changes in port selection – i.e. the action switching behavior. We find that mouse behavior deviates from the theoretically optimal agent performing Bayesian inference in a hidden Markov model (HMM). Instead the strategy of mice can be well-described by a set of models that we demonstrate are mathematically equivalent: a logistic regression, drift diffusion model, and 'sticky' Bayesian model. Here we show that switching behavior of mice is characterized by several components that are conserved across models, namely a stochastic action policy, a representation of action value, and a tendency to repeat actions despite incoming evidence. When fit to mouse behavior, the expected reward under these models lies near a plateau of the value landscape even in changing reward probability contexts. These results indicate that mouse behavior reaches near-maximal performance with reduced action switching and can be described by models with a small number of relatively fixed-parameters.

Introduction

Animals must select appropriate actions to achieve their goals. Furthermore, animals must adapt their decision-making process as the environment changes. During foraging, for example,

animals must make decisions about when and where to search to safely acquire sufficient nutrients. This requires balancing the tradeoff between exploiting known sources of food while continuing to explore unknown, potentially more profitable options. In a dynamic environment, continued exploration and adaptation are required to detect and react to changing conditions, such as the depletion or appearance of a food source, that may influence the optimal decision at a given time.

The dynamic multi-armed bandit task is an experimental paradigm used to investigate analogs of these decision-making behaviors in a laboratory setting (Samejima et al. 2005; Daw et al. 2006; Tai et al. 2012; Parker et al. 2016; Ebitz, Albarran, and Moore 2018; Hattori et al. 2019; Bari et al. 2019; Donahue, Liu, and Kreitzer 2018; Costa, Mitz, and Averbeck 2019). In this task, animals choose between a small number of actions, each of which offers a nonstationary probability of reward. In this paradigm the dynamic reward contingencies require the players to flexibly modulate their actions in response to evidence accumulated over multiple trials. Therefore, switching between behaviors is a key component of performing this task. However, analysis of behavior in this task is often reduced to examining the animal's selection of the higher rewarding port, and trials in which the animal switches between actions are not explicitly considered.

The neural mechanisms underlying the decision-making strategy employed by animals in the multi-armed bandit are poorly understood, but are thought to involve basal ganglia and medial prefrontal cortex, including key inputs from neuromodulatory systems such as dopaminergic neurons (Ebitz, Albarran, and Moore 2018; Verharen, Adan, and Vanderschuren 2019; Bari et al. 2019; Gershman and Uchida 2019). Recent work has shown the suitability of logistic regression and reinforcement learning-based models in predicting the choice behavior of agents, providing insight into how simple algorithms can reduce a series of actions and outcomes to features that are neurally tractable (Ito and Doya 2009; Miller, Botvinick, and Brody 2018). These models of behavior have facilitated the identification of neural correlates of action value representations, as well as neural activity corresponding to exploration in which the expressed behavior deviates from the action with the highest expected value (Tai et al. 2012; Ebitz, Albarran, and Moore 2018; Donahue, Liu, and Kreitzer 2018; Verharen, Adan, and Vanderschuren 2019).

Here, we develop a statistical analysis of the relatively infrequent subset of trials in which the agent switches between actions, enabling examination of the features that contribute to the flexible and exploratory components of behavior. We use these models to study mouse behavior in a two-armed bandit task and gain insight into the strategy that animals use to select actions to achieve reward. We find that trial-to-trial action switching is a stochastic component of the behavior and sets theoretical limits on the performance of behavioral models in predicting action choice. Although the optimal agent in this task would perform inference in a hidden Markov model (HMM), mouse behavior is not consistent with that of such an agent. Instead it is better-described by a simple logistic regression using a stochastic action-selection policy. By leveraging the simple form of the logistic regression weights, we formulate a drift diffusion model with a choice-history bias that not only captures mouse choice and switching behavior, but generalizes to new environmental parameters through a parsimonious solution that costs the agent minimal expected rewards. Finally, we relate this drift diffusion model to a 'sticky' hidden Markov model, yielding multiple equivalent models that capture animal behavior and make predictions about the neural mechanisms underlying the observed behavior.

Results

Task structure and performance

To study probabilistic decision making, we trained mice in a Markovian two-armed bandit task. During each behavior session, the mouse moved freely in a chamber containing three ports into which it could place its snout (i.e. nose poke) to engage with the task (Figure 1A). The center port was used for trial initiation, and the two side ports delivered fixed-sized water rewards according to preassigned reward probabilities, such that on any given trial the mouse had a high and low reward probability option. The probability of water delivery at the high probability port was 0.8 (i.e., P(reward|high-choice)=0.8) and that at the low probability port was 0.2 (i.e., P(reward|low-choice)=0.2). The state of the reward probabilities was assigned on a trial-by-trial basis following a Markovian process, such that the high and low assignments reversed with a probability of 0.02 after completion of each trial. This stochastic process produced blocks of consecutive trials during which the high reward probability was assigned to the right or left port (Figure 1B), with a mean block length of 50 trials.

Wild-type mice learned to perform this task and typically achieved an average of 514 ± 77 water rewards in a 40 min session (± SD, n=6). Overall, right and left port selection was unbiased (51% left, 49% right) and mice performed each trial quickly (center port to center port elapsed time or trial durations of mean ± SD = 2.05 ± 3.14 s or median ± MAD = 1.65 ± 0.79 s). The mean time between center and choice port was 0.47 s, much faster than the 2 s upper limit imposed by the task structure. Decision times (center port to side port) were broad, with trials in which the mouse switched between ports taking longer on average, and some mice also exhibited behavior consistent with the post-reinforcement pause (Ferster and Skinner 1957) (Supplementary Figure 1). Information about action timing was not used in the analyses and models presented below.

To quantify task performance and characterize the behavioral strategy, we considered two aspects of the mouse choice on each trial. First, we determined the probability of selecting the higher rewarding port ($P_{highchoice}$), which reflects the ability of the mouse to collect information across trials to form a model of the optimal action. Second, we measured the probability of switching port selection from one trial to the next (P_{switch}), which reflects the trial-to-trial propensity of the mouse to alter its action choices. Switch trials occurred infrequently, making up only 0.07 of all trials. Each mouse made decisions in a clearly non-random pattern: across all mice, $P_{highchoice}$ was 0.83 (range of 0.81 to 0.84 across mice) resulting in reward delivery on 0.70 (range of 0.69 to 0.70 across mice, Supplementary Table 1) of trials (compared to expected reward rate for random choices of 0.5 = 0.8*0.5 + 0.2*0.5). Furthermore, the strategy employed by the mice deviated from a simple “win-repeat, lose-switch” strategy as P_{switch} was 0.02 following rewarded choices and 0.18 following unrewarded choices (as opposed to the 0.0 and 1.0 rates predicted by win-repeat, lose-switch).

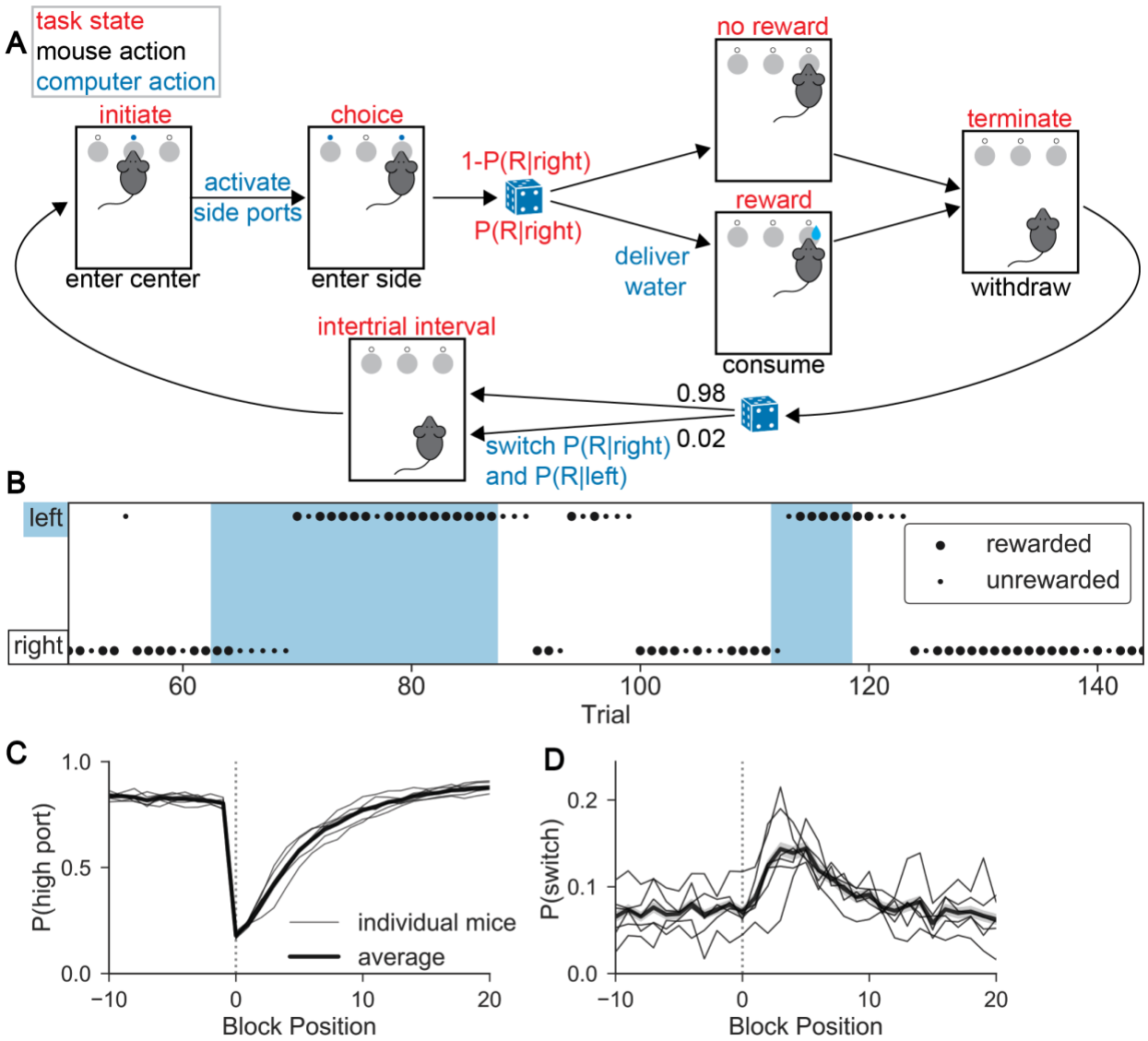


Figure 1. Mouse behavior in a two-armed bandit task. (A) Task structure: A mouse initiates a trial by putting its snout (i.e. "poking") into the center port. It then selects one of the two side ports in order to enter the "choice" state. In this illustration, the mouse chose the right port. Depending on the choice and preassigned port reward probabilities, reward is or is not delivered. The mouse "terminates" the trial by withdrawing from the side port, which initiates the "inter-trial interval" state. During this 1 s period, the computer assigns reward probabilities for the subsequent trial using a Markov process. (B) Example mouse behavior across part of a session. Blue and white shading indicates the location of the high reward probability port as left and right, respectively. Dot position and size indicate the port chosen by the mouse and the outcome of the trial, respectively (large = rewarded). (C) Summary of probabilities with which mice chose the high reward probability port (P(reward)=0.8) as a function of trial number surrounding the trial at which the reward probabilities reverse (block position=0). Each thin line shows the behavior of an individual mouse (N=6) whereas the thicker line and the shading around it show the mean and standard error, respectively, across mice. (D) As in (C) but showing the probability that mice switch port choice on trial *n* (block position) compared to their choice on trial *n-1*.

Mice were sensitive to the task structure and dynamic reward probability assignments: mice generally chose the higher rewarding port but adjusted their behavior in response to reward

probability reversals at block transitions (Figure 1C). The rate of selection of the high probability port fell steeply following the block transition as perseverance of mice on their previous port choice corresponded to selection of the lower rewarding port. The mice required multiple trials to stably select the new high reward probability port after a block transition ($\tau = 6.17 \pm 0.65$ s.e.m. trials) (Figure 1C, Supplementary Table 1). This performance was achieved through an increased switch rate immediately after the block transition: although across all trials P_{switch} was low, it increased after the block transition (Figure 1D), paralleling the recovery of $P_{highchoice}$. The dynamics of $P_{highchoice}$ and P_{switch} following the block transitions indicates that mice, as expected, modulate their behavior in response to the outcomes of choices and motivates our pursuit of models that capture this behavioral strategy (Tai et al. 2012; Parker et al. 2016; Donahue, Liu, and Kreitzer 2018; Bari et al. 2019).

History dependence of behavior

To examine the contribution of trial history to mouse choice, we computed the conditional probability that the mouse switched ports given each unique combination of choice-reward sequences in the preceding trials (P(switch|sequence)). This can be thought of as a nonparametric policy in which the combination of previous choices and rewards (implicitly across varying latent states) guides future choice (Figure 2A). To represent the conditioned history sequences, each trial was given a label that captured both action (relative choice direction) and the outcome of that action (reward or no reward): the letter (a/A vs b/B) denoted the action and the case (lower vs. upper) denoted the outcome with upper case indicating a rewarded trial. We defined the first choice direction of the sequence as “A”, so that, depending on reward outcome, choices in this direction were also labeled “A/a” whereas those in the other direction were labeled “B/b.” This code was used to build a “word” (e.g. Aab) that fully specifies port choice and action outcome over a chosen history length (3 in the given example) leading up to each trial (Figure 2B). For a history length of 3 trials, switching behavior has left-right symmetry, confirming our ability to represent direction in relative terms (Figure 2C).

The number of possible conditioned sequences expands exponentially as history length increases: there are 4 possibilities when considering a trial history of length 1 but over 10^6 for length 10. Across all sessions (~115,000 trials) the observed mouse behavior contained a large

number of unique sequences (e.g. >21,000 unique sequences of length 10). However, the rapid expansion of possible sequences prevents full analysis of the contribution of histories beyond length 3 or 4 trials, as evidenced by the proportion of sequences that occur with a standard error greater than, for example, 20% (Supplementary Figure 2). Previous analyses of similar tasks (Tai et al. 2012; Hattori et al. 2019; Belkaid et al. 2020) found the contributions of trial history beyond length 3 or 4 was negligible. Based on cross-validated likelihood estimates on held-out data (Supplementary Figure 2), we will present here analysis of conditional switch probability using history length 3. Nevertheless, we included in the supplementary figures an extension to longer history lengths by restricting analysis to sequences that are sufficiently represented in the behavioral data, which are consistent with the results described for history length 3 (Supplementary Figure 2).

Stochasticity of behavior limits the single trial accuracy of predictive models

To characterize the history dependence of the mouse switching behavior, we examined conditional switch probabilities for all unique action and outcome sequences for history length 3 (Figure 2D). Two notable features emerge from this analysis. First, the probability of switching varies as a function of trial history, confirming that mouse behavior depends on action and outcome history. Broad trends can be identified such as the tendency to repeat the previous action after rewarded trials. Second, although mice exhibit a regime of behavior in which they nearly deterministically repeat the same port choice on subsequent trials (P_{switch}~0), the maximum conditional P_{switch} does not approach 1 for any action/outcome sequence, instead reaching a maximum of ~0.5 (P(switch | “Abb”) =0.47 ± 0.078 s.e.m.). Thus, switches cannot be predicted with certainty for any combination of three past actions and outcomes. This apparent stochasticity persists for longer history sequences that are expressed sufficiently often to reliably calculate P(switch | sequence): for history lengths 4 and 5, the maximum conditional P_{switch} among sequences with standard error below 20% were P(switch | “Abbb”)=0.54 ± 0.13 and P(switch | “aAAAb”)=0.58 ± 0.19, respectively (Supplementary Figure 2). Thus, mouse behavior can, in this framework, be qualitatively described as moving from an “exploit” state of repeating recently rewarded actions to an “explore” state of random port choice after recent failures to receive reward.

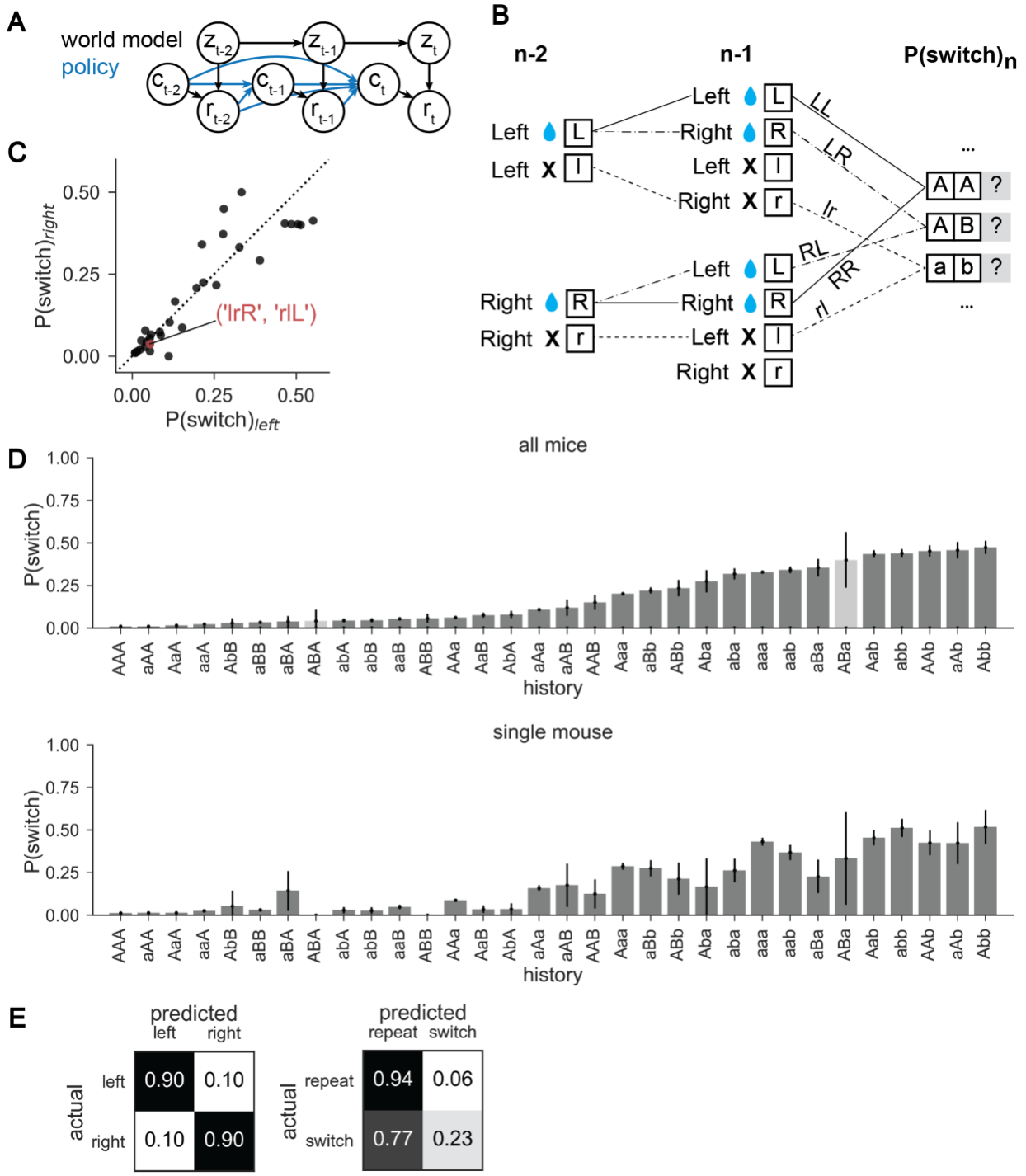


Figure 2. Switching behavior is probabilistic and history dependent. (A) Schematic of world model (black lines) for the two-armed bandit task: rewards (r) depend on mouse choice (c) and the underlying state (z) for each trial (t). World state evolves according to a Markov process. A nonparametric policy (blue) shows previous choices and rewards contributing to future choice. (B) The action-outcome combination for each trial is fully specified by one of four symbols: "L" or "R" for left or right rewarded trial, "l" or "r" for left or right unrewarded trial, respectively. These can form "words" that represent action-outcome combinations across sequences of trials. Each sequence starting with right port selection has a mirror sequence starting with a left port selection (e.g. r-L and l-R, in panel A) and can be combined by defining the initial direction in the sequence as "A/a." The probability of switching

ports on the next trial is calculated, conditioned on each trial sequence for history length *n*. (C) The conditional switch probabilities after R/L mirror pairs of history length 3 are plotted for histories starting on the left vs. right port. The clustering of points around the unity line confirms the symmetry of mouse switching (correlation coefficient = 0.91). One such pair (l-r-R and r-l-L) is highlighted, which becomes a single sequence (a-b-B). (D) *top:* Conditional switch probability across all mice for each action-outcome trial sequence of history length 3, sorted by switch probability. Each bar height indicates the mean switch probability following the corresponding action-outcome history across all trials and mice. The error bars show binomial standard errors. Sequences that occur with s.e.m.>20% are shown in lighter gray. *bottom:* As above for data collected across all sessions for a single representative mouse. Sequences are presented along the x-axis using the same order as in the top graph. (E) Confusion matrices for the nonparametric policy for right and left port choice (*left*) and repeat and switch (*right*). On-diagonal values represent the theoretical maximum for sensitivity, or the proportion of predicted positives relative to all positives, under the mouse's conditional probability distribution. Off-diagonal values represent expected proportion of false negatives, normalized to one across the row with true positives.

For a history of length 3, this nonparametric model of mouse behavior is defined by $\frac{4^3}{2} = 32$ conditional probabilities. A more concise summary is given by its confusion matrices, the average probability it assigns to the mouse's choices (Figure 2E). We considered two representations of these choices: the chosen port (left/right) and whether the mouse switched port from the last trial (repeat/switch). These confusion matrices show that left and right port choice are highly predictable actions, each with an average probability of 0.90. In contrast, although the repetition of action selection from one trial to the next is highly predictable, with an average probability of 0.94, the stochastic nature of switching events makes them highly unpredictable, such that the probability of predicting the mouse will switch its port choice from one trial to the next is only 0.23. Nevertheless, this prediction is better than that expected by chance given the 0.06 basal switch rate.

To maximize the overall likelihood of the mouse's behavior, a model should calibrate its predictions to match the conditional probability of the next choice, just like the nonparametric model. We use confusion matrices like the ones in Figure 2E to evaluate the subsequent models, which incorporate constraints to shed light on the algorithms by which mice make decisions. A good model should assign high likelihood to the mouse's choices, capture the stochasticity of behavior in its conditional probabilities of switching, and exhibit similar confusion matrices as the nonparametric model.

Models of mouse behavior

Our goal in the analysis of mouse behavior presented above was to extract features that could be quantified and used to constrain and test models of behavior. Based on this analysis we selected

three criteria to evaluate models of mouse behavior, alongside canonical model comparison with held-out log likelihood:

1) The ability of the model to accurately predict port selection and switching events on a trial-by-trial basis, as compared to the expected confusion matrices defined above (Figure 2E, Methods).

2) The ability of the model to capture the conditional action and outcome history dependence of P_{switch} (Figure 2D).

3) The ability of the model to reproduce the dynamics of $P_{highchoice}$ and P_{switch} around block transitions (Figure 1C-D).

These features of behavior were stable within and across sessions (Supplementary Figures 3,4).

We separately consider two components underlying the observed behavior: algorithm and policy, in which the former is the process used to generate 'beliefs' about the state of the environment (i.e. level of confidence that the higher reward port is left vs. right), and the latter relates those computed beliefs to a decision to select a port. Due to the stochasticity of mouse conditional switch probabilities, we hypothesized that a stochastic action policy is needed to best describe the observed behaviors. Therefore, in testing models that perform deterministic computations, we expect that "greedy" type policies in which the action associated with the higher probability of reward is always selected will perform less well than stochastic policies in which some subset of actions are selected randomly.

Hidden Markov Model as the ideal agent fails to capture mouse behavior

The behavioral task was designed to evolve according to a discrete Markovian process and, therefore, from the agent's perspective the world can be described as governed by a hidden Markov model. In our two-armed bandit task, there are two environmental states, z_t (i.e., corresponding to states in which either the left or right port is the higher reward probability port) that are not directly observable by the mouse. Instead, these states are relayed to the mouse through the state emission probabilities – the probability that the mouse receives reward given the current state and its port selection, $P(r_t|z_t, c_t)$. Thus, an optimal agent uses Bayesian inference in a hidden Markov model (HMM) to estimate its belief in the environmental state (see Methods): computing and recursively updating its posterior belief (b_t) in the latent state by

incorporating incoming evidence from sequences of mouse actions and reward outcomes (Figure 3A, Methods):

$$b_{t+1} = \sum_{k} P\left(z_{t+1} | z_t = k\right) \cdot P(r_t | z_t = k, c_t) \cdot b_{t,k}$$

In which c_t is the mouse's choice and r_t is the reward outcome for trial t. This belief estimate is then passed through an action policy to predict the mouse's choice on the next trial.

To test if an HMM could accurately model the mouse's behavior, we used a grid search over all parameters and selected the model that maximized the log likelihood of the data given the model. The best-fit model accurately represents the temporal structure of the environment (maximized at a transition probability of 0.02) but underestimates the high port reward probability (maximized at an emission probability of 0.65 whereas the true probability was 0.8). We predicted mouse behavior using a stochastic probability matching policy, in which the port favored by the model is selected at a rate proportional to the model's belief (also known as Thompson sampling) (Thompson 1933).

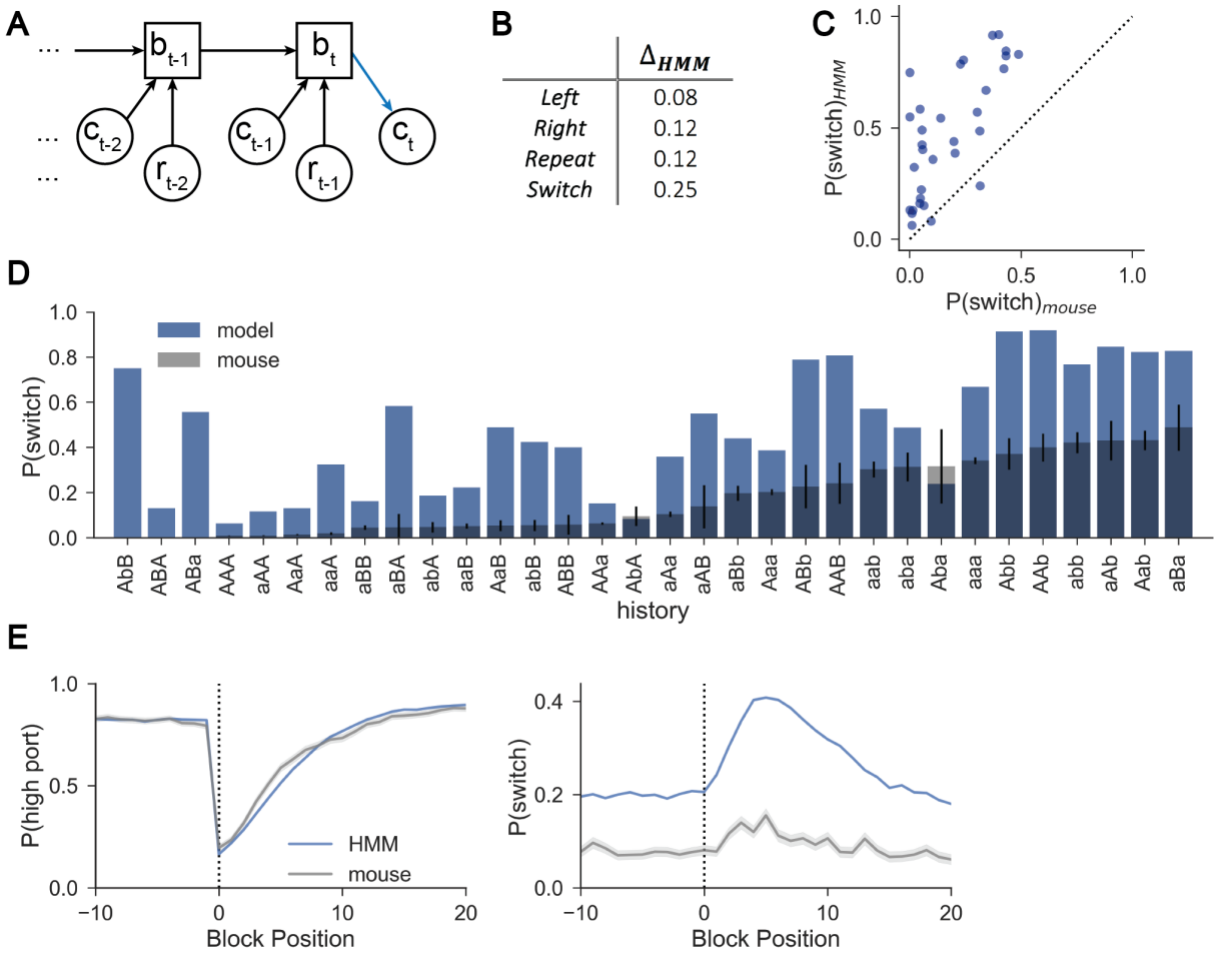


Figure 3. Hidden Markov model overpredicts mouse switching behavior. (A) The Hidden Markov model (HMM) recursively updates belief state (b_{t+1}) by incorporating evidence from choice (c_t) and reward (r_t) of the recent trial. The next choice (c_{t+1}) depends on the model posterior and the policy (blue). (B) Absolute values of the differences between the HMM confusion matrices and nonparametric confusion matrix (Figure 2E) for each action type. (C) Conditional switch probabilities generated from the HMM plotted against those observed from mice (SSE = 4.102). (D) Conditional switch probabilities as predicted by the HMM (blue, 'model') overlaid on the observed mouse behavior (gray) for all history sequences of length 3. Sequences on the x-axis are sorted by increasing $P(\text{switch})_{mouse}$ as in Figure 2D. The bar heights show the mean switch probability across mice for each corresponding sequence history, and the error bars show the binomial standard error for the mouse test data. (E) HMM-generated probability (blue) of choosing the high reward probability port (*left*) and of switching ports (*right*) as a function of trial number surrounding state transition (block position 0) as compared to the mouse behavior (gray). Dark lines show the mean across trials at the same block position and the shading shows the standard error.

Thompson sampling in an HMM fails to capture essential features of the mouse behavior by systematically overpredicting the probability of switching (Figure 3B-E). This is reflected by the deviation of the model from the expected confusion matrices of the nonparametric policy, which we compute as the absolute values of the differences between the model's values and expected values for each action (Figure 3B Δs, compared to the data in Figure 2E). Accordingly, the

model overpredicts the conditional switch probabilities. (Figure 3C-D). (Note: Here we present the analyses of the held-out data not used for training, which is only 30% of the data presented in Figure 2D. For this reason, the orderings of history sequences by mouse conditional switch probabilities and the binomial standard error estimates differ across figures.) Finally, the HMM fails to capture the dynamics of $P_{highchoice}$ and P_{switch} around block transitions of reward probabilities (Figure 3E). We also examined the HMM parameters that correspond to the ideal observer, acting under both a Thompson sampling and greedy policy (i.e. one that deterministically selects the port that has a higher probability according to the model's belief), but on each of the behavioral features outlined above, these models also failed to capture the mouse behavior (Supplementary Figure 5).

Logistic regression with a stochastic policy better predicts mouse behavior

We next considered logistic regression, which has been used previously to describe rodent behavior in similar tasks (Tai et al. 2012; Parker et al. 2016; Donahue, Liu, and Kreitzer 2018; Miller, Botvinick, and Brody 2018), as an alternative model. Although this simpler model was shown to perform well at predicting the right and left choice of animals in these tasks, its ability to predict switches has not been evaluated. We built a logistic regression that computed the log-odds of the mouse's next choice as a function of past choices and rewards,

$$\psi_{t+1} = \sum_{i=0}^{L_1} \alpha_i \bar{c}_{t-i} + \sum_{i=0}^{L_2} \beta_i \bar{c}_{t-i} r_{t-i} + \sum_{i=0}^{L_3} \gamma_i r_{t-i} + \epsilon$$

where $\alpha, \beta,$ and γ represent the weights on input features for choice ($\bar{c}_t$), encoding of choice-reward interaction ($\bar{c}_t r_t$), and reward (r_t) across trials back to L_1, L_2, and L_3, respectively. In this logistic model the choice is represented by $\bar{c}_t$, in which -1 indicates a right port choice and 1 a left port choice, and the reward output is represented by r_t, in which 0 indicates no reward and 1 reward. We fit the model and used cross validation to select the number of past trials to include for each feature. This confirmed that there is minimal left-right port choice bias (i.e., ϵ = 0.04). We also found that rewards alone did not contribute significantly to choice prediction (i.e., L_3 = 0) but that the history of choice-reward encoded trials benefited the model (i.e., L_2=5, Figure

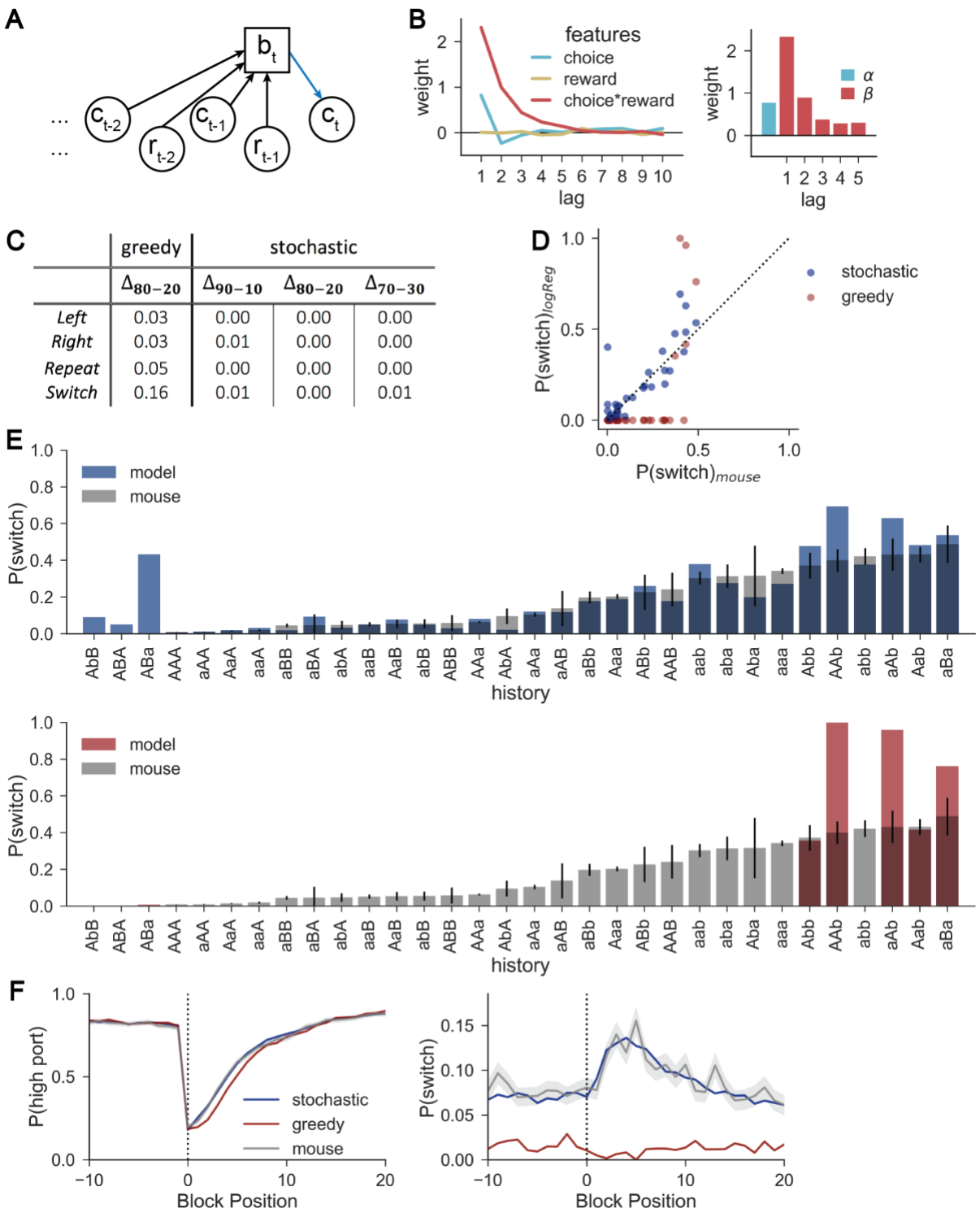
4B). Furthermore, only information about the most recent port choice was necessary (i.e., L_1=1). This enabled us to use a reduced form of the model log-odds computation:

$$\psi_{t+1} = \alpha \bar{c}_t + \sum_{i=0}^{5} \beta_i \bar{c}_{t-i} r_{t-i}$$

The feature weights indicate a propensity of mice to repeat their previous action, as denoted by the positive coefficient on previous choice (hereby denoted by α, Figure 4B).

We tested the fit model on the remaining (held-out) data to predict the left or right choice of the mouse and found that this model coupled with a stochastic action policy recapitulated all features of the behavior and achieved comparable log-likelihood estimates on held-out data to those of the nonparametric model (Figure 4C-F, blue traces; Supplementary Table 2). The stochastic policy used here, and in all models below, selects a port at a rate proportional to the model estimate (see Methods). The stochastic logistic regression captured both the port choice and switching behavior of the mouse as well as possible given the stochastic nature of behavior (i.e., $\Delta \sim 0$, Figure 4C). This is further supported by the conditional switch probabilities predicted by the model as compared to the mouse (Figure 4E), in which it is evident the model captures both the history dependence of the mouse's switching behavior, as well as the stochastic nature with which it applies a decision-making policy. Finally, the model recapitulates the time course over which the block transition perturbs stable port selection and uses increased switch prediction as a mechanism to recover the selection of the high port (Figure 4F).

These results differ from those of the Bayesian model (i.e. HMM) as well as from the same logistic regression model using a deterministic policy (greedy policy maximizing log-odds choice from the model; Figure 4, red traces). Interestingly, the impact of policy on model performance is most evident when evaluating model fit on switching behavior, with surprisingly subtle effects on the model's accuracy in predicting left vs. right choice (Figure 4F). Although the greedy logistic regression captures much of the dynamics of $P_{highchoice}$ (Figure 4F, left), it does so without predicting switching between ports (Figure 4F, right). These results emphasize the need for explicit examination of switch trials in behavioral modeling.



	greedy	stochastic		
	Δ_{80-20}	Δ_{90-10}	Δ_{80-20}	Δ_{70-30}
Left	0.03	0.00	0.00	0.00
Right	0.03	0.01	0.00	0.00
Repeat	0.05	0.00	0.00	0.00
Switch	0.16	0.01	0.00	0.01

Figure 4. Stochastic logistic regression policy captures mouse behavior comprehensively, whereas greedy logistic regression fails to predict switches. (A) The logistic regression computes the probability of choice (b_{t+1}) from choice (c_t) and reward (r_t) information across a series of trials. Here we represent the model estimate as b for consistency across graphical representations, but note that it in this case it corresponds to the log-odds of choice, ψ, in the text. (B) *left:* Feature weights for a logistic regression predicting the log-odds of mouse port selection for the choices, rewards, and the choice-reward interactions in the previous 10 trials. *right:* Feature weights after cross-

validation for hyperparameters and refitting the model. α is the weight on the previous choice, and β is the set of weights on choice-reward information for the previous 5 trials. (C) Absolute value of the differences between the logistic regression confusion matrices and nonparametric confusion matrix (Figure 2E) for each action. Δ scores are shown for stochastic logistic regression across three sets of probability conditions, as well as for greedy logistic regression in the P(reward|high-choice)=0.8 condition. (D) Conditional switch probabilities generated by the logistic regression model using a stochastic (blue) or greedy (red) policy plotted against those observed in mice (stochastic-SSE = 0.378, greedy-SSE = 1.548). (E) *top:* Conditional switch probabilities for the stochastic logistic regression (blue) across sequences of history length 3 overlaid on those from the mouse data (gray). Sequences on the x-axis are sorted according to mouse conditional switch probabilities. Error bars show binomial standard errors for the mouse. *bottom:* As above but for a greedy policy (red). (F) Probabilities of choosing the high reward probability port (*left*) and of switching ports (*right*) as a function of trial number surrounding state transition (block position 0). Logistic regression predictions with a stochastic (blue) and greedy (red) policy are overlaid on probabilities observed for mice (gray). Dark lines show the mean across trials at the same block position and the shading shows the standard error.

Reduced logistic regression captures behavior in different reward probability conditions

To determine whether the reduced form of the logistic regression model generalizes to other environmental conditions, we tested the same mice on two new sets of reward probabilities (P(reward | high-choice)=0.9 and P(reward | low-choice)=0.1, P(reward | high-choice)=0.7 and P(reward | low-choice)=0.3). We found that the mice modified their behavior in the different task conditions (Supplementary Figure 6). Nevertheless, the same model structure fit to the data obtained under these new probability conditions produced predictions that matched the values of the expected confusion matrices across the four action types (Figure 4C), demonstrating the generalizability of this model.

Drift diffusion formulation of the reduced logistic regression

Our goal in modeling behavior was to uncover the task features and algorithms that lead to the expressed decision-making strategy. Although a reduced logistic regression can accurately capture the mouse behavior, it provides an inefficient neural solution by requiring the weights on each of these features to be learned and the past choices and rewards to be stored in memory. We therefore asked how such an algorithm could be approximated by the animal and learned over time.

We inspected the structure of the logistic regression model to determine whether we could achieve similar predictive accuracies with a recursively updated algorithm. We found that the weights assigned to past choices and rewards were well fit by an exponential curve, with initial magnitude β that decays across trials at a rate of τ (Figure 5B). Plugging in the exponential

approximation, and approximating the finite sum with an infinite one (since $\tau < L_2$), we can rewrite the log-odds of port selection on the next trial (ψ_{t+1}) as,

$$\psi_{t+1} \approx \alpha \bar{c}_t + \beta \sum_{i=0}^{\infty} e^{-i/\tau} \bar{c}_{t-i} r_{t-i}$$

Furthermore, we can define this exponential term as the recursive quantity ϕ_t,

$$\phi_t \triangleq \beta \sum_{j=0}^{\infty} e^{-j/\tau} \bar{c}_{t-j} r_{t-j}$$

$$= \beta \bar{c}_t r_t + e^{-j/\tau} \cdot \beta \sum_{j=0}^{\infty} e^{-j/\tau} \bar{c}_{t-1-j} r_{t-1-j}$$

$$= \beta \bar{c}_t r_t + e^{-1/\tau} \phi_{t-1}.$$

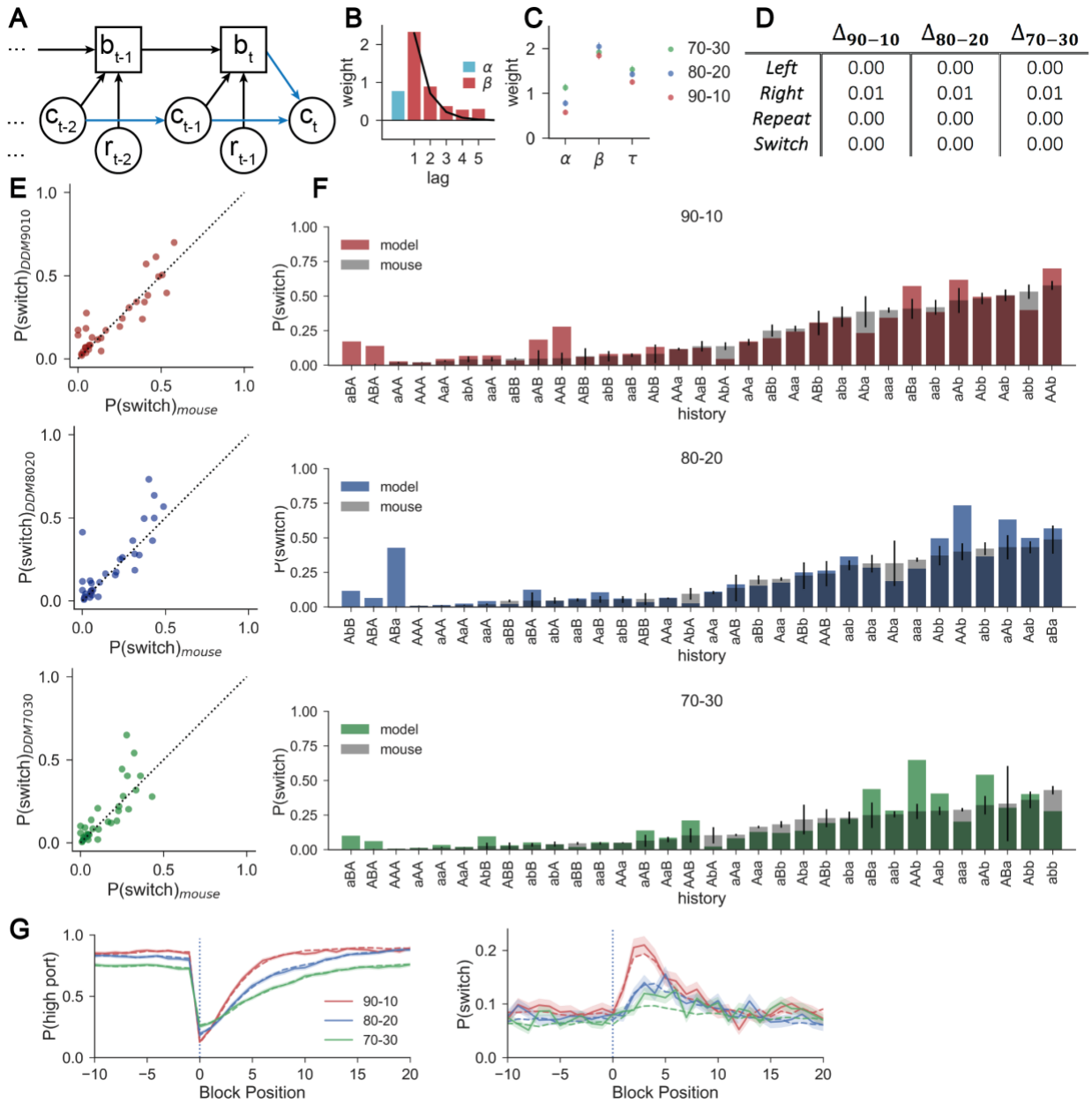
We recognize the resulting form as a type of drift diffusion model (DDM) (Ratcliff and McKoon 2008; Pedersen, Frank, and Biele 2017; Urai et al. 2019) that decays toward zero with time constant τ, but receives additive inputs depending on the most recent choice and whether or not it yielded a reward. The magnitude β determines the weight given to incoming evidence. Therefore, our computation of the log-odds can be given as a filtering of choices and rewards biased by α toward the most recent choice (Figure 5A):

$$\psi_{t+1} = \alpha \bar{c}_t + \phi_t.$$

This form of the model offers two advantages over the original logistic regression when considering a potential neural implementation of the algorithm: 1) the exponential representation of choice and reward history captures the behavior using a model with only three free parameters (α, β, τ), whereas the logistic regression used six, and 2) the recursive definition of this choice-reward representation reduces the memory demands of the model.

We tested the drift diffusion model on all three task variants (P(reward|high-choice)∈{0.9,0.8,0.7}), and found it predicted all features of mouse behavior excellently (Figure 5D-G, Supplementary Table 2). Interestingly, we found that the α parameter varied the most across reward probability conditions, while β and τ remained relatively constant (Figure 5C), suggesting that the mechanism by which mice adapted their behavior can be explained by

increasing or decreasing their bias toward repeating their previous choice. Notably, α>0 in all contexts, such that there was always some tendency to repeat the previous choice ("stickiness").



	Δ_{90-10}	Δ_{80-20}	Δ_{70-30}
Left	0.00	0.00	0.00
Right	0.01	0.01	0.01
Repeat	0.00	0.00	0.00
Switch	0.00	0.00	0.00

Figure 5. A drift diffusion model formulation of the logistic regression recapitulates behavior in multiple reward probability conditions. (A) A drift diffusion model with choice history bias recursively updates a single state belief (b_{t+1}) using evidence from recent choice (c_t) and reward (r_t). The policy (blue) shows an additional contribution on next choice prediction from the previous choice. (B) β weights for choice-reward information are described by an exponential function (black curve). (C) Summary of the fit DDM parameters for data from mice performing in the three sets of reward probability conditions. Each data point shows the mean parameter estimate with error bars indicating the bootstrapped 95% confidence intervals. (D) Δ scores for absolute values of the differences between the DDM confusion matrices and nonparametric confusion matrix (Figure 2E) for each action

across the three reward probability conditions. (E) Conditional switch probabilities calculated from the DDM predictions plotted against those of the observed mouse behavior for each set of reward probability conditions (*top*: 90-10 (SSE=0.243), *middle*: 80-20 (SSE=0.417), *bottom*: 70-30 (SSE=0.33)). (F) Conditional switch probabilities predicted by the DDM (model) across sequences of history length 3 overlaid on those from the mouse data (gray) for the three sets of reward probability conditions. Error bars show binomial standard error for the mouse. (G) Probabilities of choosing the high reward probability port (*left*) and of switching ports (*right*) as a function of trial number surrounding state transition (block position 0) for the three sets of probability conditions. Dashed lines show mean of model predictions, solid lines show mean of true mouse probabilities across trials at the same block position. Shading shows the standard error.

Returning to the HMM

The drift diffusion model derived from the empirical logistic regression resembles the optimal Bayesian model, but with an additional influence of previous choice on future choice (compare graphical representations in Figures 3A and 5A). To gain insight into the differences between these models, we developed a mathematical correspondence for the log-odds computation by the DDM with that of the HMM (see Methods), allowing direct comparison of parameter estimates. This revealed that the deviation of the mouse from the optimal player (HMM) can be explained by the constraints on α, the stickiness or bias towards repeating the last action, in the HMM. Whereas for all reward probability conditions the fit with a DDM yields $\alpha>0$ (Figure 5C), the optimal HMM requires $\alpha<0$. This difference can be conceptualized as a sticky tendency in the empirical model that biases the mouse to repeat its previous action, in contrast to the HMM, which makes its selection considering only its posterior belief and independent of any additional choice history. Phrased in a different way, the HMM retains no information about the identity of the last port choice but instead actively uses evidence of a reward on the last trial to update its prior belief about the identity of the highly rewarded port.

To explicitly capture the concept of "port switching", we incorporated a second state variable into the HMM to capture the influence of recent choice history on future choice. We model this as a dynamic cost for the mouse of switching actions as the difference in log-odds between the HMM's current belief state and that of the DDM (see Methods). Including this additional term makes the computed trial-by-trial log-odds of the HMM the same as those of the DDM, ensuring that both capture all features of the behavior equally well.

Comparison of behavior of models performing the two-armed bandit task

Analysis of the trial-by-trial log-odds estimates for the DDM (and accordingly for the sticky HMM) reveal asymmetrical use of rewarded vs. unrewarded choice information, whereby rewarded choices provide evidence toward the selected port, but unrewarded choices result in a decay toward α (and therefore maintaining a slight preference for the most recent choice, Figure 6A). This asymmetry has been previously reported in analysis of mouse evidence accumulation (Vertechi et al. 2019). This contrasts the mechanics of the optimal agent, where unrewarded trials provide evidence toward the alternative port (Figure 6A). For an optimal agent, an unrewarded choice (or series of unrewarded choices) at the current selection port can flip the sign of belief or log-odds ratio, providing evidence in favor of switching ports (P_{switch}>0.5, and even nearing deterministic P_{switch}) in conflict with the actual mouse behavior. In other words, a switch can be caused by a change in the belief of which port has the higher reward probability.

In contrast, for the empirically better-fitting models (logistic regression, DDM, and sticky HMM), the effect of unrewarded trials on the log-odds estimate is to drift towards its choice history bias (i.e. α) and, therefore, like the mouse, cause increasingly random port selection. Shifting the port favored by the empirical models requires achieving a reward on the alternative port from the current preference, which causes an update and sign flip in the belief parameter. This suggests that switches under the empirical models rely on the combination of the odds ratio approaching 1 (i.e. log-odds=0) and a stochastic action policy to facilitate random sampling of the low-probability port. It is these stochastic switches – rather than evidence-based switching – that allows the model to update its belief to favor a new action in the future. In rare cases, unrewarded switches flip the sign of α, and so potentially shift the log-odds in favor of the new port (depending on the magnitude of ϕ_t), but this behavior is consistent with the necessity for a stochastic switch to precede evidence favoring the alternative port over the previous port.

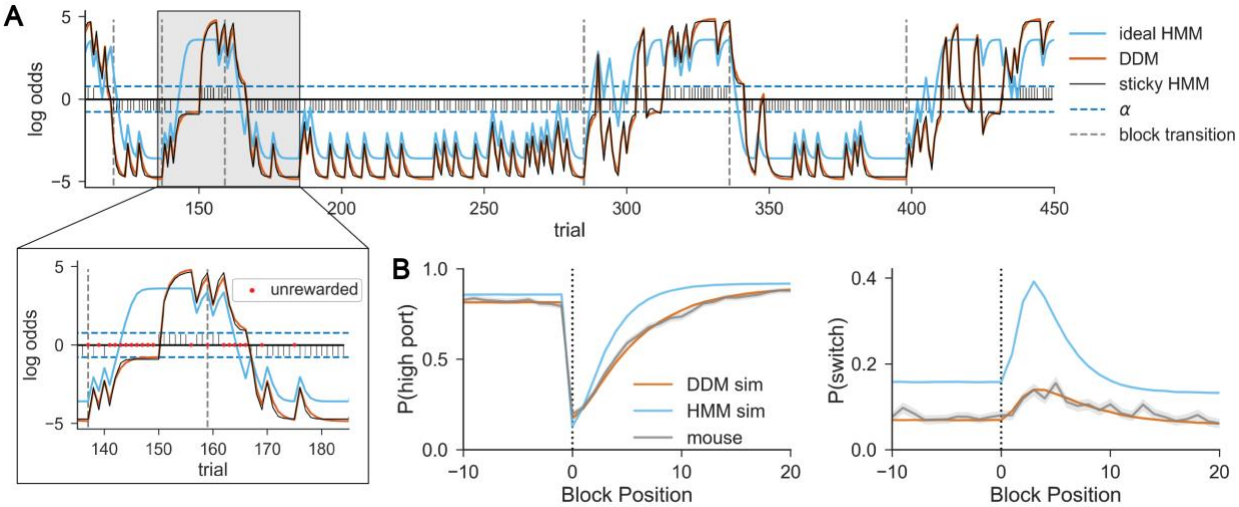


Figure 6. Simulations with a generative drift diffusion model recapitulate mouse behavior. (A) Representative session depicting equivalent trial-by-trial log-odds computations for the DDM vs. the sticky HMM (orange vs. black traces). These model estimates contrast the log-odds of the posterior computed by the ideal HMM (light blue), which specifically diverges in prediction updating following unrewarded trials. Stem plot shows the choice*reward interaction that provides action-outcome evidence to the DDM. Horizontal dashed lines indicate ±α, and vertical dashed lines indicate state transitions. *Inset*: Expanded segment of session with unrewarded trials labeled by red dots. (B) Probabilities of choosing the high reward probability port (*left*) and of switching ports (*right*) as a function of trial number surrounding a state transition (block position 0) in the P(reward|high-choice)=0.8 context for the generative DDM (orange) and generative ideal HMM (light blue) overlaid with the observed mouse probabilities (grey). The lines show the means across trials at the same block position and the shadings show the standard errors.

Comparison of dynamics of the model algorithms also reveals that they exhibit different bounds on the maximum and minimum trial-to-trial switching probability (see Methods; Figure 6A). The upper and lower bounds of switching probability in the ideal HMM are constrained by the odds ratio of the transition probability – the model's log-odds belief in the identity of high reward probability port is bounded by the probability that the port reward probability stays the same from trial-to-trial. In contrast, the DDM and sticky HMM reach steady-state near-deterministic behavior (Figure 6A, Methods). These bounds explain the elevated switch rate produced by Thompson sampling on the HMM belief state, even outside of the block transition. Following reward, the belief log-odds of the HMM are further constrained to the product of the odds ratios of the emission probability and transition probability.

The deviation of the empirical behavior from the theoretically optimal model appears striking when considering history-dependent action selection. However, it is unclear that these deviations have a significant cost in terms of the total rewards received. Surprisingly, the expected reward rate of the original Thompson sampling HMM predicting choice from mouse behavior was only

marginally better than that actually achieved by the mice (71% vs. 70% trials rewarded in sessions where P(reward|high-choice)=0.8, respectively). To determine whether this was an effect of the suboptimality of the mouse history crippling the HMM performance, we simulated data under the ideal HMM unbounded from mouse history. We initialized an HMM with the true task parameters (P(reward|high-choice)=0.8 and $P(z_{t+1} \neq z_t)=0.02$) and allowed it to play the game using its own past choices and rewards as history. This model did not perform better, achieving 71% ± 0.4% rewards per session (mean ± s.e.m.). We compared this performance to simulations run under a generative form of the DDM using the empirically fit parameters, which achieved 69% ± 0.0% rewards per session (mean ± s.e.m.). Notably, even without the mouse history as input features to guide action selection, the DDM behavior resembles the characteristic patterns of mouse behavior (Figure 6B).

We hypothesized that the mice converged to a local maximum or plateau of expected reward within the parameter space in which further optimization of behavior driven by reward rate is challenging. For each of the three reward contexts we held τ constant at the corresponding empirically fit value and examined expected reward across the two-dimensional parameter space for varying α and β. In each, there is a wide plateau over which expected reward stabilizes, and both the α and β values for the true task parameters under the original HMM and the fit values under the DDM lie near this plateau (Figure 7A). For this reason, near maximal performance can be achieved with a broad range of α and β values (Figure 7B).

We also considered that the mice may optimize reward relative to a cognitive or physical cost, as opposite to optimizing reward rate at any cost. Specifically, we hypothesized that the stickiness of the empirical models might indicate a preference for the mice not just for reward maximization, but efficient collection of reward in terms of behavioral effort, in this case as reflected in the switching rate. Comparing the ratio of rewards to switches, we found that the DDM achieves twice as many rewards per switch as the Thompson sampling HMM in the P(reward|high-choice)=0.8 condition (i.e. an average of 9.95 vs. 4.46 rewards/switch, respectively). Calculating this ratio of rewards per switch for models simulating behavior in each reward context, we find that the DDM exceeds the original HMM in all three (Figure 7C). Interestingly, this parallels minimal differences between the models in overall expected reward

(Figure 7C), and so can be attributed to the DDM's efficient reduction in switching. However, both of these models are outperformed by a greedy HMM, suggesting that the DDM's advantage to maximizing rewards per switch depends on first selecting a stochastic policy. This suggests that, under the assumption that switching ports bears a cognitive and/or physical cost, and given a tendency for exploration, the objective of the mice is not exclusively reward maximization, but rather optimizing the tradeoff between reward maximization and cost.

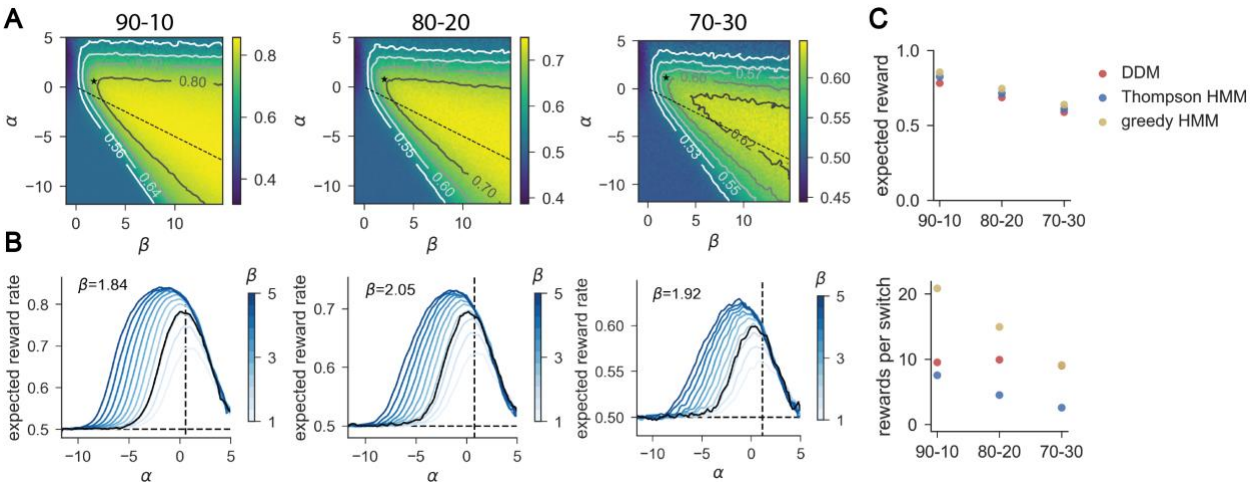


Figure 7. Reward per switch ratios differentiate models and policies that all achieve near-maximal expected reward. (A) Expected reward landscape for the generative DDM across varying α (y-axis) and β (x-axis) values with the empirically observed τ in each of the three reward contexts (τ_{90-10} =1.25, τ_{80-20}=1.43, τ_{70-30}=1.54). Color bars indicate expected reward rate across simulated trials, and isoclines mark increments above random (0.5). The DDM-fit α and β values are depicted at the asterisk (*), and the relative α and β specified for the HMM lie along the dashed line. (B) Profile of expected reward as a function of α for varying values of β (color bar, ranging from β=1 to β=5, with fit β in black). Expected reward rate at the fit α (black vertical dashed line) suggests minimal additional benefit of modulating β. (C) *top*: Expected reward in each of the three probability contexts for the generative DDM using mouse-fit parameters and generative HMM using the true task parameters. HMM performance is shown using either a greedy or stochastic (Thompson sampling) policy. *bottom*: Ratio of rewards to switches for each of the three models across contexts. Each data point shows the mean across simulated sessions and error bars show standard error but are smaller than the symbol size.

Discussion

Switch trials reveal stochasticity in mouse behavior

Many behavioral tasks, including the two-armed bandit, are described as having components of "explore vs. exploit" in which an agent at times exploits existing knowledge and executes an action most likely to lead to reward, whereas at other times it explores the environment by choosing an action with a less certain outcome that reveals information about the environment

(Daw et al. 2006; Dayan and Daw 2008; Costa, Mitz, and Averbeck 2019; Hattori et al. 2019; Pisupati et al. 2021; Rosenberg et al. 2021).

Despite the richness offered by the analysis of behavioral dynamics and of deviations from "optimal" strategies, the description of mouse behavior in the two-armed bandit task is often reduced to simply stating the agent's propensity to select the higher rewarding port ($P_{highchoice}$) across the behavioral session. Similarly, the performance of models of behavior in the two-armed bandit task is typically evaluated by their ability to predict port selection across the session. This often reduces to examining the model's ability to identify the highly rewarded port, as the majority of trials occur in extended stretches of steady-state exploit-like behavior in which the mouse repeatedly selects the more highly rewarded port.

In contrast, few studies explicitly describe or model trials in which the animal switches its action choice from that expressed in the previous trial. Since these switch trials occur infrequently, the overall performance of models of behavior in a two-armed bandit task is relatively unaffected by categorically failing to accurately predict their occurrence. However, in a two-armed bandit task, the trials in which the agent switches ports are the manifestation of behavioral flexibility (i.e., changes in action due to accumulating information) and exploration, and are highly informative components of the behavior. Analysis of these trials provides an important insight by revealing the stochastic nature of mouse decision making. Although mice enter a regime of nearly deterministic repetition of actions (i.e., during exploit phases), they do not enter a corresponding regime of deterministic switching (i.e., no accumulation of evidence will consistently push the mouse to switch actions). Thus, even following a series of 'no reward' outcomes at a single port, the mouse will choose its next action apparently at random rather than reliably switch selection to the other port.

This understanding propels our selection of a stochastic policy to best represent the action policy of the mice, which captures the tendency of the mice to make decisions at a rate proportional to their confidence in those decisions. The stochastic action policy evident during behavior in the two-armed bandit task balances trials in which mice exploit information (i.e., favors the port with the higher probability of providing a reward, given the trial history) and explore alternatives (i.e.

deviates from this prediction) (Hattori et al. 2019; Vertechi et al. 2019). It should be noted that the stochasticity we describe is observed under the constraints of our model variables and history length, but does not necessarily characterize the decision to switch given an unconstrained model (i.e. given a complete history or access to neural activity). Clearly, at the extreme, the exact sequence of actions and action outcomes expressed by the mouse leading up to a trial late in a session is likely unique (given the exponential growth in sequence possibilities as a function of trial number), and thus it is not possible to determine if the action choice is stochastic given the full history.

Stochasticity of behavior constrains maximum predictability of behavior by models

There has been a recent push in behavioral studies to account for behavioral events at the resolution of single trials (Williams and Linderman 2021). This is a worthwhile goal, especially in evaluating the predictive performance of behavioral models. However, we found that the stochastic component of behavior creates a tradeoff between model accuracy at the single trial level and across the full distribution of trials. Therefore, we compared the performance of each model against the theoretical probabilities of predicting each action (i.e., expected confusion matrices from the nonparametric model) set by the stochasticity of the mouse behavior on the same type of trial. This proved a powerful approach, allowing us to evaluate our models in the context of the constraints imposed by the inherent stochastic nature of the behavior. In the context of exploratory behavior, the method described here or a similar approach to constraining models under the true distribution of the data (Rosenberg et al. 2021), enables testing of models against realistic boundaries of predictive accuracy.

Stickiness captures the deviation of mouse behavior from optimality

Interestingly, although perhaps not surprisingly, we find that the model that best recapitulates the mouse behavior does not use the algorithm that maximizes reward in this task (the HMM). Single latent variable HMMs can be implemented in artificial neural networks and therefore at least in principle by the brain, so it is unclear why mice do not perform this optimal strategy (Tran et al. 2016). An ethological explanation can be proposed from our observation that using the optimal strategy offers only marginal increases in expected reward over another simple computational algorithm (i.e., the drift diffusion model derived from the logistic regression)

(Roy et al. 2021). Moreover, given a tendency for exploration or stochasticity, the HMM requires more trials in which the agent switches between ports to achieve equal reward. This hypothesis suggests that constraints imposed by learning the task and task structure or asymmetric costs associated with the selection or executions of actions lead the mouse away from the HMM implementation. Additionally, it brings up an interesting question as to whether mice have an innate tendency for exploration in environments with uncertainty (Grunow and Neuringer 2002; Tervo et al. 2014; Belkaid et al. 2020; Lai and Gershman 2021).

The differences between the HMM and the empirical logistic regression (drift diffusion model with choice history bias) can be captured by an additional influence of past choice on future choice. We account for this by building a sticky HMM, which, by construction, produces equivalent trial-by-trial log-odds predictions as the DDM. Therefore, we find two models with distinct mechanics that equivalently recapitulate mouse behavior, and can thus examine conserved aspects of the two models to make hypotheses about features necessary to generate the observed behavioral strategy. The first feature we have discussed extensively, namely the necessity for a stochastic action policy on the model estimate. Secondly, both models encode the interaction between choice and reward rather than the variables independently, consistent with previous accounts of action value encoding in brain regions such as striatum (Samejima et al. 2005; Kim et al. 2009; Tai et al. 2012; Donahue, Liu, and Kreitzer 2018). Both models use a state representation to efficiently store the memory of the computed estimate, which is recursively updated with incoming evidence on each trial. Lastly, as discussed above, both models require “stickiness” in action choice.

This stickiness has been reported in analyses of behavior across tasks and species, and is also called perseveration, choice history bias, and the law of exercise (Thorndike 1911; Ito and Doya 2009; Balcarras et al. 2016; Miller, Botvinick, and Brody 2018; Urai et al. 2019; Lak et al. 2020; Gershman 2020; Lai and Gershman 2021). We find that this bias to repeat previous actions offers a parsimonious mechanism for adapting an existing action policy to novel environmental conditions. When we changed the reward probability conditions in our two-armed bandit task, we found that mice minimally updated the weights on incoming evidence and memory decay (β and τ, respectively), but instead modulated their behavior by increasing or decreasing their level

of perseveration. This behavioral adaptation, represented largely by a single parameter, comes at low cost to the animal in terms of expected reward, and therefore may be an efficient strategy for minimizing effort necessary to learn new behavioral strategies (Fan, Gold, and Ding 2018; Drugowitsch et al. 2019; Gershman 2020).

Implications for the neural mechanism

One of the goals of this study was to increase our understanding of decision making in order to guide future interrogation of circuit function and the neural underpinnings of behavior. However, the specific algorithms that we found best fit the mouse behavior may or may not be directly implemented in the brain. The demonstration that multiple distinct algorithms can similarly model behavior underscores this point, and draws our focus in considering neural representations to the features described above that are shared by the models. We hypothesize that whatever algorithm the brain relies on for this task, it is combined with a stochastic action policy to produce the behavior we observe. We note that it is possible that a policy that appears stochastic behaviorally can be traced to neural origins that are deterministic.

Furthermore, past work has hypothesized that recursive algorithms that compress information over a sequence of trials to a small number of variables are neurally plausible (Dayan and Daw 2008). Here, we show that some recursive algorithms (i.e., original HMM) struggle to explain switching behavior, while non-recursive models (i.e., logistic regression) perform well. This poses a potential challenge to this hypothesis. However, we were to able derive alternative recursive algorithms (i.e., DDM and sticky HMM) that explain behavior. Additional models that capture more complex state representations beyond the latent structure of the task will likely be important in parsing neural activity that corresponds to nonstationary decision policies, as described in recent work (Ashwood, Roy, and Bak 2020; Zoltowski, Pillow, and Linderman 2020).

<u>Materials and Methods</u>

Behavior apparatus

The arena for the two-armed bandit task was inspired by previous work (Tai et al. 2012). Behavior experiments were conducted in 4.9” x 6” custom acrylic chambers. Each chamber contained three nose ports with an infrared-beam sensor (Digi-Key, 365-1769-ND) to detect entry of the snout into the port. A colored LED was positioned above each port. For the two side ports, water was delivered in 2.5 µL increments via stainless steel tubes controlled by solenoids (The Lee Co, LHQA0531220H). The timing of task events was controlled by a microcontroller (Arduino) and custom software (MATLAB). Plans for an updated version of the behavioral system, including the most recent hardware and software, are available online: https://edspace.american.edu/openbehavior/project/2abt/ and https://github.com/bernardosabatinilab/two-armed-bandit-task

Behavior task

Wild-type mice (*C56BL/6N* from Charles River and bred in house) aged 6-10 weeks were water restricted to 1-2 mL per day prior to training and maintained at >80% of full body weight. While performing the task, mice moved freely in the chamber. Activation of an LED above the center port indicated that the mouse could initiate a trial by nose poking into the center port. Doing so activated LEDs above the two side ports, prompting the mouse to choose to nose poke to the right or left. The mouse had 2 s to make its selection. Following side port entry, the computer determined whether or not to deliver a water reward according to the corresponding port reward probability and the result of pseudo-random number generation. Withdrawal from the side port ended the trial and started an inter-trial interval (ITI). The 1 s ITI followed selection, during which time the system assigned the reward probabilities for the next trial according to a Markov decision process (0.98 probability that high and low port assignments remained the same, 0.02 probability the assignments reversed). After the 1 s minimum ITI, the center port LED turned on and the mouse was permitted to initiate the next trial (with no upper limit to trial initiation time). The duration of each behavior session was 40 minutes, over which the mouse typically earned >350 rewards. All training sessions were conducted in the dark or under red-light conditions.

Data Analysis

Models of mouse behavior

All behavior models were trained on 70% of sessions and tested on the remaining held-out data. For models predicting mouse choice on previous mouse behavior (Figures 3-5), model predictions were taken as the mean across 1000 repetitions on bootstrapped test data to acquire representative estimates of choice and switch probabilities.

Hidden Markov model of mouse port choice behavior

We built an HMM that computes its posterior belief (b_{t+1}) that on the next trial t the latent state of the system (z_{t+1}), i.e., the port most likely to yield reward, is equal to k. This belief state is recursively updated by incorporating evidence from the previous trial, and used as a predictive estimate by incorporating knowledge of the transition matrix. To compute the probability b_{t+1}, we take the sum across the joint probability of transitioning from state $z = k$ on the next trial and the likelihood of observing the specific action-outcome combination given that state (emission probability), weighted by the prior belief for both right and left states (the exhaustive set of states):

$$b_{t+1} = \sum_k P\,(z_{t+1}|z_t = k) \cdot P(r_t|z_t = k, c_t) \cdot b_{t,k}.$$

Here, as a single order Markovian model, the probability of a state transition is fully captured by the current state so that $P(z_{t+1}|z_t) = P(z_{t+1}|z_{1:t})$, where z_t represents the state at trial t and $z_{1:t}$ all the states up to the current one. The likelihood of evidence from the most recent trial (emission probability) is captured by the conditional probability of reward (r), given the latent state (z) and the mouse's choice (c).

We implemented the ideal HMM using Python code available on GitHub (https://github.com/lindermanlab/ssm). For the ideal agent we built a model with two discrete latent states and access to the true transition and emission matrices. For example, for the version of the task where P(reward|high-choice)=0.8, the emission probabilities $P(r_t|z_t, c_t) = [0.8, 0.2]$ and transition probabilities $P(z_{t+1}|z_t) = [0.98, 0.02]$. To test the HMM's performance at predicting mouse choice, we fed the model action and outcome data from sequences of trials. For each session, the model was initialized with equal priors for the right and left state, after which the model iteratively updated its belief by advancing through the trial sequence. b_{t+1} has upper

and lower bounds constrained by the nonstationary dynamics of the latent state, captured by the predictive probability of $P(z_{t+1}|z_t)$. The posterior estimate for each trial was passed through an action policy to make a prediction of mouse behavior (see below).

Hidden Markov model fit to mouse behavior

We considered that the world-state transition and emission probabilities hard-coded into the HMM are not known to the mouse. To optimize the fit of our HMM to the mouse's behavior, we ran a grid search over transition and emission probabilities on bootstrapped training data and calculated the log likelihood of the data given the model using each parameter set. We maximized this function to select the parameters used in our "fit HMM" and evaluated on bootstrapped test dataset (30% of data) as presented.

Logistic regression

We compute the conditional probability of choice given data from previous choices and rewards using a logistic regression to compute the log-odds, ψ:

$$\mathrm{P}(c_{t+1} = 1|\mathrm{c}_{1:\mathrm{t}}, \mathrm{r}_{1:\mathrm{t}}) = \sigma(\psi_{\mathrm{t}+1})$$

The full logistic regression model uses the weighted linear combination of the action (choice), reward outcome, and action-outcome interaction history from previous trials to calculate the log-odds of mouse choice for the next trial:

$$\psi_{\mathrm{t}+1} = \sum_{i=0}^{L_1} \alpha_i \bar{c}_{t-i} + \sum_{i=0}^{L_2} \beta_i \bar{c}_{t-i} r_{t-i} + \sum_{i=0}^{L_3} \gamma_i r_{t-i} + \epsilon,$$

where $\alpha, \beta,$ and γ represent the weights on input features for choice, encoding of choice-reward interaction, and reward across trials back to L_1, L_2, and L_3, respectively. For the features, $\bar{c}_{t-i}$ represents whether the mouse made a choice to the left or right (1 or -1 respectively), r_{t-i} represents whether or not the mouse received a reward (1 or 0, respectively), and $\bar{c}_{t-i} r_{t-i}$ the interaction between choice and reward (1 when rewarded left, -1 when rewarded right, 0 otherwise) on the *i*th-back trial. ϵ represents the overall port bias. To fit the model, we split our data into training and testing datasets. We used cross validation to fit the hyperparameters L_1, L_2,

and L_3 as 1, 5, and 0 respectively, to arrive at the reduced form of the model presented in the text and Figure 4:

$$\psi_{t+1} = \alpha \bar{c}_t + \sum_{i=0}^{5} \beta_i \bar{c}_{t-i} r_{t-i}$$

Formulation of a drift diffusion model from empirical logistic regression

We use the exponential function,

$$\beta_i = \beta e^{-\frac{i}{\tau}},$$

to approximate the weights on the encoded choice-reward history. Substituting this approximation in the reduced logistic regression, we compute the log-odds as:

$$\psi_{t+1} \approx \alpha \bar{c}_t + \beta \sum_{i=0}^{\infty} e^{-i/\tau} \bar{c}_{t-i} r_{t-i}$$

from which we define the recursive quantity, ϕ_t,

$$\phi_t \triangleq \beta \sum_{j=0}^{\infty} e^{-j/\tau} \bar{c}_{t-j} r_{t-j}$$

$$= \beta \bar{c}_t r_t + e^{-j/\tau} \cdot \beta \sum_{j=0}^{\infty} e^{-j/\tau} \bar{c}_{t-1-j} r_{t-1-j}$$

$$= \beta \bar{c}_t r_t + e^{-j/\tau} \phi_{t-1}.$$

Thus, we define the log-odds as:

$$\psi_{t+1} = \alpha \bar{c}_t + \phi_t.$$

We fit the free parameters $\alpha, \beta,$ and τ using stochastic gradient descent on the training set and estimated parameter error with bootstrapped confidence intervals.

Mathematical correspondence between HMM and DDM

The HMM is characterized by the belief state $b_{t+1} = \left(b_{t+1,L}, b_{t+1,R}\right)$ representing the distribution over the latent state z_{t+1} given preceding choices and rewards. Since the belief state is a probability vector, we know $b_{t+1,R} = 1 - b_{t+1,L}$. As described above, the belief states can be computed recursively to account for information obtained on each trial. Here we show how the recursive belief state updates are related to the recursive log-odds calculations of the DDM.

To make the correspondence, rewrite the belief state updates in terms of the log-odds ratios. Let $\Psi_{t+1} = \sigma^{-1}\left(b_{t+1,L}\right) = \log\frac{b_{t+1,L}}{1-b_{t+1,L}}$ denote the log-odds ratio of the belief state at trial $t+1$. Like the belief state, the log-odds can be computed recursively,

$$\Psi_{t+1} = f\left(\Psi_t^{(c)}; q\right) \text{ where } \Psi_t^{(c)} = \sigma^{-1}(p)\,\bar{c}_t\bar{r}_t + \Psi_t$$

and where

$$f\left(\Psi_t^{(c)}; q\right) = -\sigma^{-1}(\mathrm{q}) + \log\frac{1 + e^{\Psi_t^{(c)} + \sigma^{-1}(q)}}{1 + e^{\Psi_t^{(c)} - \sigma^{-1}(q)}} \approx \sigma^{-1}(q)\tanh\left(\frac{2q-1}{\sigma^{-1}(q)}\Psi_t^{(c)}\right)$$

In these equations, in a manner analogous to the definition of $\bar{c}_t$, we use $\bar{r}_t$ which takes the value 1 if a reward was delivered and -1 if no reward was received. $\bar{r}_t$ can be calculated from r_t (which was used above and takes values of 0 and 1 for unrewarded and rewarded outcomes, respectively) as $\bar{r}_t = 2r - 1$. In addition, $q \in [0.5, 1)$ is the probability that the system state remains the same, p is the probability of receiving a reward upon choosing the correct port, and $\sigma^{-1}(x) = \log\frac{x}{1-x}$ denotes the "logit" function; i.e. the inverse of the logistic function. (Due to the symmetric design of the experiment, p is also the probability of *not* receiving a reward upon choosing the *incorrect* port.)

Though these equations may look rather complicated, they have many intuitive properties. First, the log-odds recursions split into a "conditioning" step in which current log-odds Ψ_t are updated with new information from the current choice $\bar{c}_t$ and reward $\bar{r}_t$ to obtain $\Psi_t^{(c)}$. This step depends on the same features as the logistic regression model presented above, and the coefficients are functions of the reward probability p.

Second, the "prediction" step passes the conditioned log-odds through a nonlinear transformation $f(\Psi_t^{(c)}; q)$ to obtain log-odds for the next trial. This nonlinear function saturates at $\pm\sigma^{-1}(q)$, implying that the log-odds cannot exceed the log-odds of the transition probability. This makes sense since, even if the mouse knew the state at trial t, there is always probability $1 - q$ that it will change on the next trial. Moreover, the nonlinearity is steepest when there is substantial uncertainty ($\Psi_t^{(c)} \approx 0$). In that regime, a rewarded choice has a large influence, whereas when the mouse is already quite certain, one more rewarded choice won't change the log-odds by much.

An important difference from the empirically determined logistic regression model is that here the model treats rewarded and unrewarded actions symmetrically – i.e. repeating an action and receiving a reward updates the log-odds ratio in the same way as switching actions and not receiving a reward (and similarly for repeat-no reward and switch-reward). In reality, the mice are less sensitive to omitted rewards with repeated actions and repeat themselves much more than the HMM predicts, as shown in the main text. This discrepancy suggests a simple modification of the ideal HMM model to account for the *stickiness* of observed behavior.

Augmenting the HMM with "sticky" dependencies

We proposed a hybrid model that combines the belief states of the HMM with the stickiness of the logistic regression model. The log-odds are given by,

$$\Pr(\, c_{t+1} = 1 \mid c_{1:t}, r_{1:t}\,) = \sigma\Big(\sum_{i=0}^{L_1} \alpha_i \bar{c}_{t-i} + \beta\, \Psi_{t+1}\Big)$$

in which Ψ_{t+1} are the log-odds of the belief state from the HMM, which are a function of the past choices and rewards, as well as the reward probability p and the transition probability q. In this case, the previous actions are explicitly given to the model, allowing a bias towards repeating the last action if α>0. This model achieves the same performance as the logistic regression model with $L_1 = 3$ preceding choices. Note that the mouse does not have access to the true rewards and transition probabilities, and the best model of mouse behavior may be obtained with different values of those parameters. Indeed, we find that the best model of mouse behavior uses an *overestimate* of the reward probability (0.91 when the true probability is 0.80) and an

underestimate of the transition probability (0.86 when the true probability is 0.98). This suggests that the mice tend to attribute randomness of rewards to changing world state.

Simplified HMM log-odds update

We can further simplify the log-odds calculation by the HMM to build intuition for the model behavior. From above, if we start with the predictive log-odds as:

$$\Psi_{t+1} = -\sigma^{-1}(\mathrm{q}) + \log\frac{1 + e^{\Psi_t^{(c)} + \sigma^{-1}(q)}}{1 + e^{\Psi_t^{(c)} - \sigma^{-1}(q)}},$$

and we define the odds ratio as:

$$R = e^{\Psi},$$

then taking the logit function, σ^{-1}, as defined above such that:

$$e^{\sigma^{-1}} = \frac{q}{1-q},$$

we can write:

$$e^{\Psi_t^{(c)} + \sigma^{-1}(q)} = R_t^{(c)}\left(\frac{q}{1-q}\right)$$

and

$$e^{\Psi_t^{(c)} - \sigma^{-1}(q)} = R_t^{(c)}\left(\frac{1-q}{q}\right).$$

Plugging this into our predictive log-odds from above, we obtained a simplified formula for the predictive odds ratio for the next trial:

$$R_{t+1} = \left(\frac{1-q}{q}\right)\left(\frac{1 + e^{\Psi_t^{(c)} + \sigma^{-1}(q)}}{1 + e^{\Psi_t^{(c)} - \sigma^{-1}(q)}}\right) = \left(\frac{1-q}{q}\right)\left(\frac{1 + R_t^{(c)}\left(\frac{q}{1-q}\right)}{1 + R_t^{(c)}\left(\frac{1-q}{q}\right)}\right) = \frac{1 - q + R_t^{(c)} q}{q + R_t^{(c)}(1-q)}$$

which is a function of the conditional odds ratio and the state transition probability q. We can define the odds that the state stays the same, as:

$$s = \frac{q}{1-q}$$

such that,

$$R_{t+1} = \frac{1 + sR_t^{(c)}}{s + R_t^{(c)}}$$

and

$$R_{t+1} = \frac{1 + sx^{\bar{c}_t \bar{r}_t} R_t}{s + x^{\bar{c}_t \bar{r}_t} R_t}$$

for which we defining the odds of the emission probabilities as:

$$x = \frac{p}{1 - p}$$

We know that the conditional log-odds of port choice are a recursive function of previous choices and rewards:

$$\Psi_t^{(c+1)} = \sigma^{-1}(p)\, \bar{c}_t \bar{r}_t + \Psi_t$$

such that we arrive at a simplified form of the conditional odds ratio:

$$R_t^{(c)} = e^{\Psi_t^{(c)}} = x^{\bar{c}_t \bar{r}_t} R_t.$$

Therefore, given that $\bar{c}_t \bar{r}_t$ is either +1 or -1, we need only consider two cases by which the odds ratio from the previous trial is updated using this current evidence, the odds ratio that the world state has changed, and the emission probabilities.

In the +1 case:

$$R_{t+1} = \frac{1 + sxR_t}{s + xR_t}$$

and in the -1 case:

$$R_{t+1} = \frac{1 + sx^{-1}R_t}{s + x^{-1}R_t} = \frac{x + sR_t}{sx + R_t}$$

Describing the update in terms of these variables provides intuition for the behavior of the ideal HMM in several key ways. First, the upper bound of the predictive odds ratio is the odds that the state stays the same, $s = \frac{q}{1-q}$, and the lower bound is the odds that the state changes, s^{-1}. Additionally, after receiving a reward, the odds ratio is further constrained by the values of the emission probabilities, such that $\frac{1}{sx} \leq R_{t+1} \leq sx$. Finally, both states are equally likely under the HMM when the conditional odds ratio, $R_t^{(c)}$, and accordingly the predictive odds ratio, R_{t+1}, are equal to 1.

Generative model simulations of behavior and expected reward

To evaluate model performance independent of the actual history of mouse behavior, we ran both the HMM and DDM as generative models to produce a simulated dataset of model

behavior. We simulated the task with the location of the high rewarding port (P(reward|high-choice)=0.8) determined by a Markovian process with a transition probability of 0.02, and preserved the session structure that the mice experienced, such that the number of trials in each session was drawn from a distribution based on the mouse behavior. Each model was given the same set of sessions, and played until a simulated dataset the same size as the mouse dataset was generated. We ran this simulation for 1000 repetitions to create the averaged performance presented in Figure 6.

For the HMM, we used the ideal model given the true task parameters, and after random initialization for the first choice allowed the model to recursively update its belief given its own actions and associated outcomes to guide future choices. We generated behavior from an HMM Thompson sampling on its belief to correspond with the stochastic policy of the DDM (Figure 6B, 7C) and acting greedily on its belief (Figure 7C). For the DDM, the model played using the fit parameters of the mouse in the corresponding P(reward|high-choice) task condition. The expected reward landscape was calculated by performing a parameter grid search with this simulation.

Policy implementation

We used both deterministic and stochastic action policies to predict choice on the subsequent trial from model posterior. The greedy policy was implemented such that the model deterministically selected the higher probability choice from the model estimate as its prediction of the mouse's choice:

$$c_{t+1} = argmax\ (m_{t+1}),$$

where m_{t+1} corresponds to the HMM's b_{t+1} and the logistic regression-derived probabilities $\sigma(\psi_{t+1})$, and c_{t+1} corresponds to the predicted choice. The stochastic policy we used selects choice at a rate proportional to the model posterior:

$$c_{t+1} = \begin{cases} 0 & \propto m_{t+1}(0) \\ 1 & \propto m_{t+1}(1) \end{cases}$$

<u>Acknowledgments</u>

We thank Linda Wilbrecht and members of her lab for advice in how to implement the two-armed bandit task. We thank the Harvard Medical School Research Instrumentation Core for help in designing and implementing the necessary hardware and software. We thank members of the Sabatini and Linderman labs for helpful advice and comments on the study and manuscript. We thank Julie Locantore for assistance in training mice. This work was funded by grants to BS and SL from the Brain Initiative (NINDS, U19NS113201, a.k.a Team Dope) and the Simons Collaboration on the Global Brain. Predoctoral fellowships from the NSF graduate research fellowship program supported CB and from the Canadian Institutes of Health Research supported SN.

Author contributions

SN initiated the project. CB, SN, SL, and BS designed and formalized the analysis. CB implemented the analysis, made the figures, and together with the other authors, wrote the manuscript.

References

Ashwood, Zoe C, Nicholas A Roy, and Ji Hyun Bak. 2020. "Inferring Learning Rules from Animal Decision-Making." *NeurIPS*, 27.

Balcarras, Matthew, Salva Ardid, Daniel Kaping, Stefan Everling, and Thilo Womelsdorf. 2016. "Attentional Selection Can Be Predicted by Reinforcement Learning of Task-Relevant Stimulus Features Weighted by Value-Independent Stickiness." *Journal of Cognitive Neuroscience* 28 (2): 333–49. https://doi.org/10.1162/jocn_a_00894.

Bari, Bilal A., Cooper D. Grossman, Emily E. Lubin, Adithya E. Rajagopalan, Jianna I. Cressy, and Jeremiah Y. Cohen. 2019. "Stable Representations of Decision Variables for Flexible Behavior." *Neuron* 103 (5): 922-933.e7. https://doi.org/10.1016/j.neuron.2019.06.001.

Belkaid, Marwen, Elise Bousseyrol, Romain Durand-de Cuttoli, Malou Dongelmans, Etienne K. Duranté, Tarek Ahmed Yahia, Steve Didienne, et al. 2020. "Mice Adaptively Generate Choice Variability in a Deterministic Task." *Communications Biology* 3 (1): 1–9. https://doi.org/10.1038/s42003-020-0759-x.

Costa, Vincent D., Andrew R. Mitz, and Bruno B. Averbeck. 2019. "Subcortical Substrates of Explore-Exploit Decisions in Primates." *Neuron* 103 (3): 533-545.e5. https://doi.org/10.1016/j.neuron.2019.05.017.

Daw, Nathaniel D., John P. O'Doherty, Peter Dayan, Ben Seymour, and Raymond J. Dolan. 2006. "Cortical Substrates for Exploratory Decisions in Humans." *Nature* 441 (7095): 876–79. https://doi.org/10.1038/nature04766.

Dayan, P., and N. D. Daw. 2008. "Decision Theory, Reinforcement Learning, and the Brain." *Cognitive, Affective, & Behavioral Neuroscience* 8 (4): 429–53. https://doi.org/10.3758/CABN.8.4.429.

Donahue, Christopher H, Max Liu, and Anatol C Kreitzer. 2018. "Distinct Value Encoding in Striatal Direct and Indirect Pathways during Adaptive Learning," March, 21.

Drugowitsch, Jan, André G. Mendonça, Zachary F. Mainen, and Alexandre Pouget. 2019. "Learning Optimal Decisions with Confidence." *Proceedings of the National Academy of Sciences* 116 (49): 24872–80. https://doi.org/10.1073/pnas.1906787116.

Ebitz, R. Becket, Eddy Albarran, and Tirin Moore. 2018. "Exploration Disrupts Choice-Predictive Signals and Alters Dynamics in Prefrontal Cortex." *Neuron* 97 (2): 450-461.e9. https://doi.org/10.1016/j.neuron.2017.12.007.

Fan, Yunshu, Joshua I Gold, and Long Ding. 2018. "Ongoing, Rational Calibration of Reward-Driven Perceptual Biases." Edited by Peter Latham and Richard B Ivry. *ELife* 7 (October): e36018. https://doi.org/10.7554/eLife.36018.

Ferster, C. B., and B. F. Skinner. 1957. *Schedules of Reinforcement*. Schedules of Reinforcement. East Norwalk, CT, US: Appleton-Century-Crofts. https://doi.org/10.1037/10627-000.

Gershman, Samuel J. 2020. "Origin of Perseveration in the Trade-off between Reward and Complexity." *Cognition* 204 (November): 104394. https://doi.org/10.1016/j.cognition.2020.104394.

Gershman, Samuel J., and Naoshige Uchida. 2019. "Believing in Dopamine." *Nature Reviews Neuroscience* 20 (11): 703–14. https://doi.org/10.1038/s41583-019-0220-7.

Grunow, Alicia, and Allen Neuringer. 2002. "Learning to Vary and Varying to Learn." *Psychonomic Bulletin & Review* 9 (2): 250–58. https://doi.org/10.3758/BF03196279.

Hattori, Ryoma, Bethanny Danskin, Zeljana Babic, Nicole Mlynaryk, and Takaki Komiyama. 2019. "Area-Specificity and Plasticity of History-Dependent Value Coding During Learning." *Cell* 177 (7): 1858-1872.e15. https://doi.org/10.1016/j.cell.2019.04.027.

Ito, Makoto, and Kenji Doya. 2009. "Validation of Decision-Making Models and Analysis of Decision Variables in the Rat Basal Ganglia." *Journal of Neuroscience* 29 (31): 9861–74. https://doi.org/10.1523/JNEUROSCI.6157-08.2009.

Kim, Hoseok, Jung Hoon Sul, Namjung Huh, Daeyeol Lee, and Min Whan Jung. 2009. "Role of Striatum in Updating Values of Chosen Actions." *The Journal of Neuroscience* 29 (47): 14701–12. https://doi.org/10.1523/JNEUROSCI.2728-09.2009.

Lai, Lucy, and Samuel J. Gershman. 2021. "Policy Compression: An Information Bottleneck in Action Selection." In *Psychology of Learning and Motivation*, S0079742121000049. Elsevier. https://doi.org/10.1016/bs.plm.2021.02.004.

Lak, Armin, Emily Hueske, Junya Hirokawa, Paul Masset, Torben Ott, Anne E Urai, Tobias H Donner, et al. 2020. "Reinforcement Biases Subsequent Perceptual Decisions When Confidence Is Low, a Widespread Behavioral Phenomenon." Edited by Emilio Salinas, Michael J Frank, Emilio Salinas, Carlos D Brody, and Long Ding. *ELife* 9 (April): e49834. https://doi.org/10.7554/eLife.49834.

Miller, Kevin J., Matthew M. Botvinick, and Carlos D. Brody. 2018. "From Predictive Models to Cognitive Models: Separable Behavioral Processes Underlying Reward Learning in the Rat." Preprint. Animal Behavior and Cognition. https://doi.org/10.1101/461129.

Parker, Nathan F., Courtney M. Cameron, Joshua P. Taliaferro, Junuk Lee, Jung Yoon Choi, Thomas J. Davidson, Nathaniel D. Daw, and Ilana B. Witten. 2016. "Reward and Choice Encoding in Terminals of Midbrain Dopamine Neurons Depends on Striatal Target." *Nature Neuroscience* 19 (6): 845–54. https://doi.org/10.1038/nn.4287.

Pedersen, Mads Lund, Michael J. Frank, and Guido Biele. 2017. "The Drift Diffusion Model as the Choice Rule in Reinforcement Learning." *Psychonomic Bulletin & Review* 24 (4): 1234–51. https://doi.org/10.3758/s13423-016-1199-y.

Pisupati, Sashank, Lital Chartarifsky-Lynn, Anup Khanal, and Anne K Churchland. 2021. "Lapses in Perceptual Decisions Reflect Exploration." Edited by Daeyeol Lee, Joshua I Gold, Long Ding, and Alex C Kwan. *ELife* 10 (January): e55490. https://doi.org/10.7554/eLife.55490.

Ratcliff, Roger, and Gail McKoon. 2008. "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks." *Neural Computation* 20 (4): 873–922. https://doi.org/10.1162/neco.2008.12-06-420.

Rosenberg, Matthew, Tony Zhang, Pietro Perona, and Markus Meister. 2021. "Mice in a Labyrinth: Rapid Learning, Sudden Insight, and Efficient Exploration." Preprint. Neuroscience. https://doi.org/10.1101/2021.01.14.426746.

Roy, Nicholas A., Ji Hyun Bak, Athena Akrami, Carlos D. Brody, and Jonathan W. Pillow. 2021. "Extracting the Dynamics of Behavior in Sensory Decision-Making Experiments." *Neuron* 109 (4): 597-610.e6. https://doi.org/10.1016/j.neuron.2020.12.004.

Samejima, Kazuyuki, Yasumasa Ueda, Kenji Doya, and Minoru Kimura. 2005. "Representation of Action-Specific Reward Values in the Striatum." *Science* 310 (5752): 1337–40. https://doi.org/10.1126/science.1115270.

Tai, Lung-Hao, A. Moses Lee, Nora Benavidez, Antonello Bonci, and Linda Wilbrecht. 2012. "Transient Stimulation of Distinct Subpopulations of Striatal Neurons Mimics Changes in Action Value." *Nature Neuroscience* 15 (9): 1281–89. https://doi.org/10.1038/nn.3188.

Tervo, Dougal G. R., Mikhail Proskurin, Maxim Manakov, Mayank Kabra, Alison Vollmer, Kristin Branson, and Alla Y. Karpova. 2014. "Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex." *Cell* 159 (1): 21–32. https://doi.org/10.1016/j.cell.2014.08.037.

Thompson, William R. 1933. "On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples." *Biometrika* 25 (3/4): 285–94. https://doi.org/10.2307/2332286.

Thorndike, E.L. 1911. *Animal Intelligence: Experimental Studies*.

Tran, Ke M., Yonatan Bisk, Ashish Vaswani, Daniel Marcu, and Kevin Knight. 2016. "Unsupervised Neural Hidden Markov Models." In *Proceedings of the Workshop on Structured Prediction for NLP*, 63–71. Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5907.

Urai, Anne E, Jan Willem de Gee, Konstantinos Tsetsos, and Tobias H Donner. 2019. "Choice History Biases Subsequent Evidence Accumulation." Edited by Timothy Verstynen, Barbara G Shinn-Cunningham, and Timothy Verstynen. *ELife* 8 (July): e46331. https://doi.org/10.7554/eLife.46331.

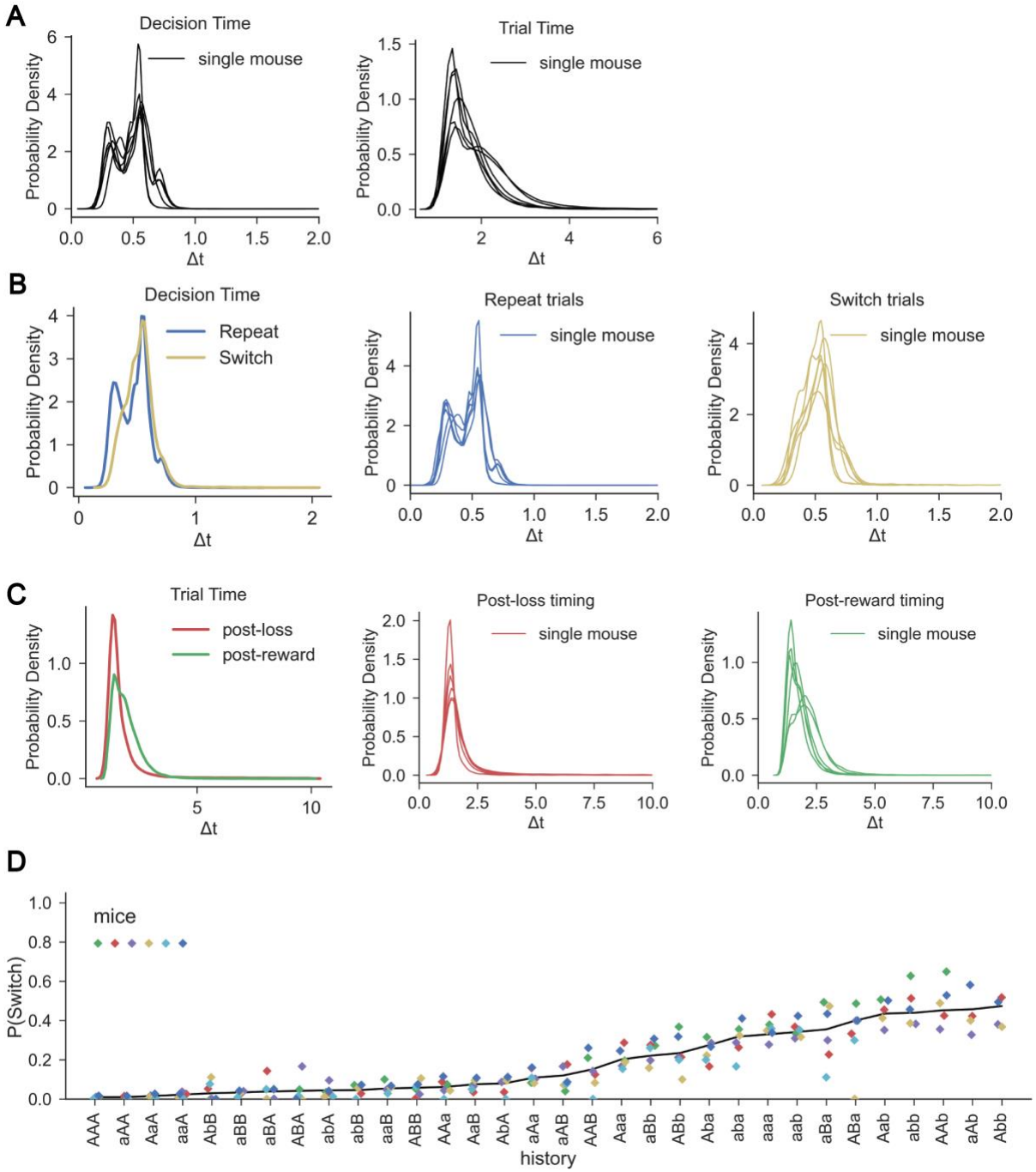
Verharen, Jeroen P. H., Roger A. H. Adan, and Louk J. M. J. Vanderschuren. 2019. "Differential Contributions of Striatal Dopamine D1 and D2 Receptors to Component Processes of Value-Based Decision Making." *Neuropsychopharmacology* 44 (13): 2195–2204. https://doi.org/10.1038/s41386-019-0454-0.

Vertechi, Pietro, Eran Lottem, Dario Sarra, Beatriz Godinho, Isaac Treves, Tiago Quendera, Matthijs Nicolai Oude Lohuis, and Zachary F. Mainen. 2019. "Inference Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas | BioRxiv." June 2019. https://www.biorxiv.org/content/10.1101/679142v1.full.

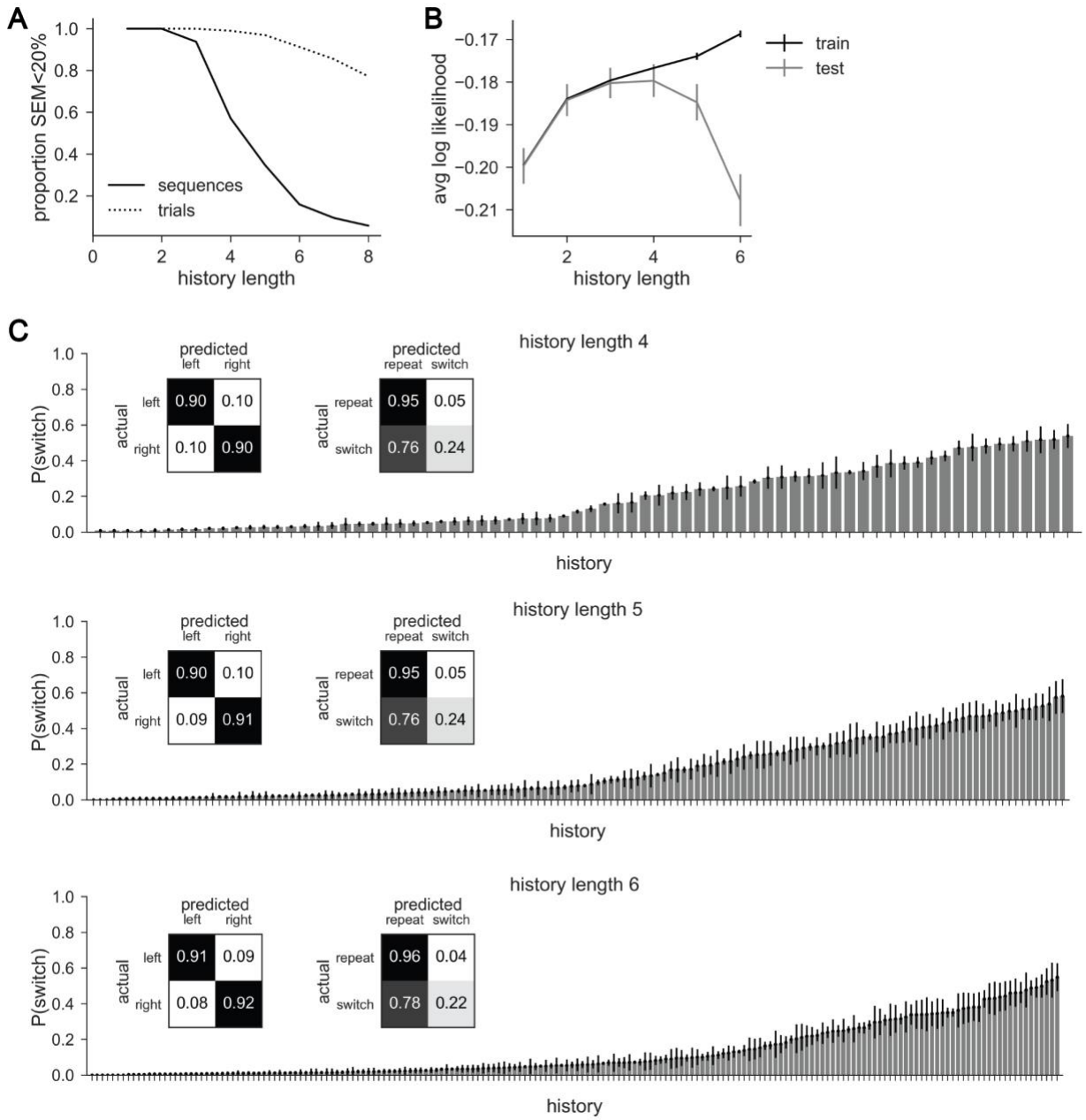
Williams, Alex H, and Scott W Linderman. 2021. "Statistical Neuroscience in the Single Trial Limit," March, 25.

Zoltowski, David M, Jonathan W Pillow, and Scott W Linderman. 2020. "A General Recurrent State Space Framework for Modeling Neural Dynamics during Decision-Making." *ICML*, 12.

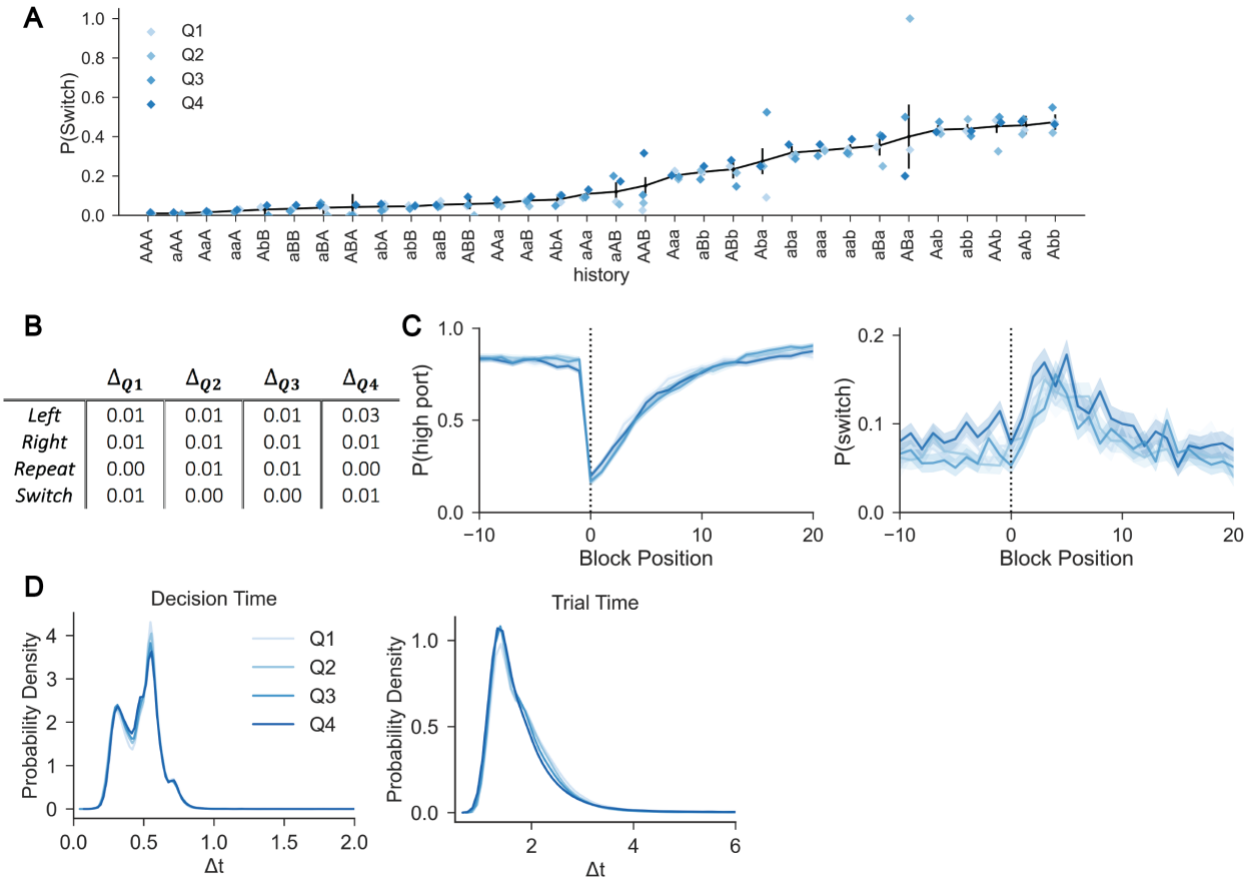
Supplementary Materials



Supplementary Figure 1. Individual mouse behavior and decision times. (A) Probability density distributions of trial times for each mouse (thin line). *left:* Distribution of time from center port to side port (decision time), *right:* distribution of time from center port to center port (trial initiation to next trial initiation). The most extreme trials ($\Delta t > 10s$, comprising <1% trials) were excluded. Probability densities integrate to 1 across the continuous distribution of durations. (B) *left:* Probability density distributions of decision times (center port to side port) for trials in which the mouse switched ports vs. those in which the mouse repeated its decision at the same port. *middle*: Distributions for individual mice on "repeat" trials, *right*: distributions for individual mice on "switch" trials. (C) *left*: Probability density distributions of trial duration (center port to center port, including intertrial interval) following reward vs. following no reward. *middle*: Distributions for individual mice following no reward, *right*: distributions for individual mice following reward. (D) Conditional switch probabilities for each action-outcome trial sequence of history length 3 for each mouse. Each symbol indicates the mean switch probability following the corresponding action-outcome history across all trials for each mouse. History sequences are sorted by the aggregate conditional switch probabilities of all mice (black line).

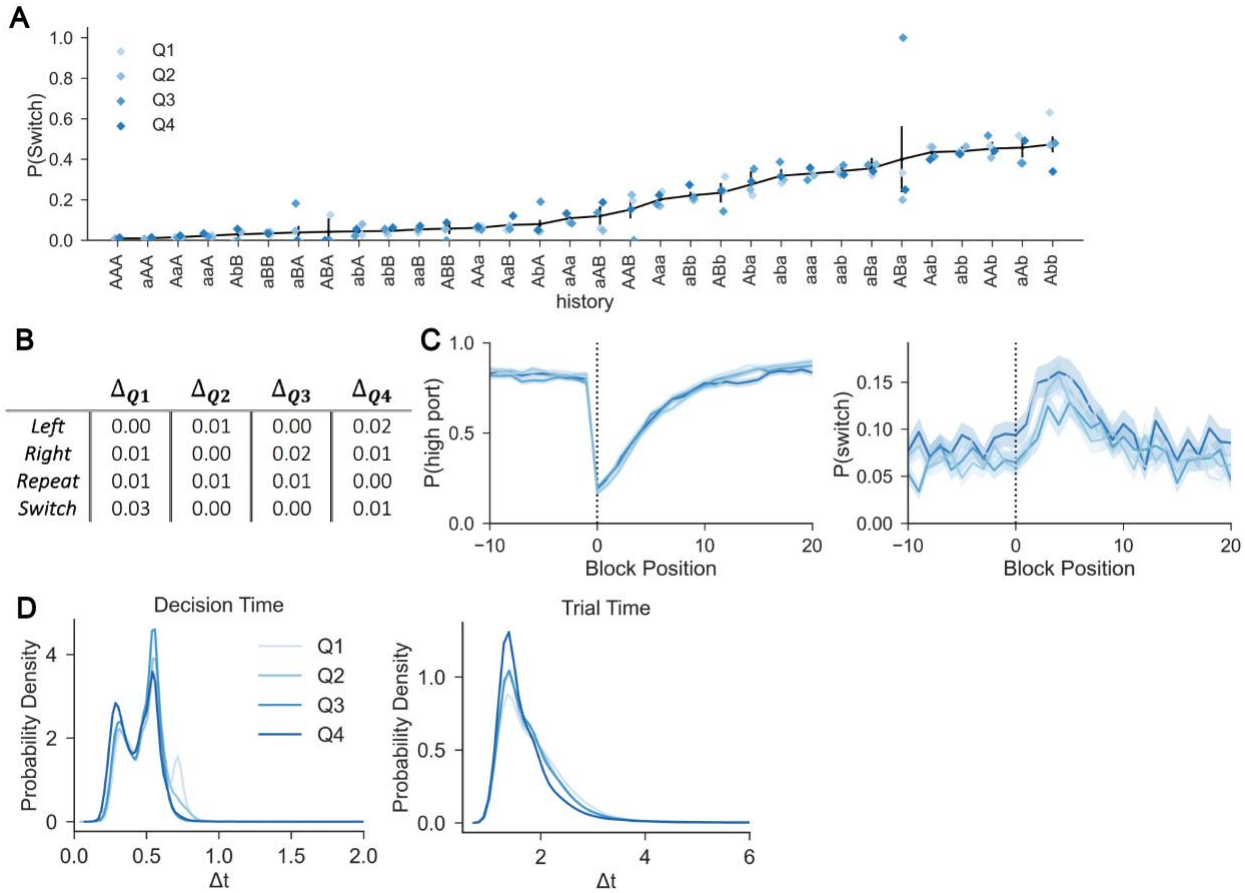


Supplementary Figure 2. Contribution of longer trial histories to conditional switch probabilities. (A) Proportion of expressed action and outcome history sequences with standard error (SEM) less than 20% (solid line), and the corresponding proportion of trials (dashed line) as a function of increasing history length. (B) Average likelihood estimates for the empirical nonparametric policies on training and testing data, using 5-fold cross validation. Error bars show standard error of the estimates. (C) Conditional switch probabilities given action-outcome trial sequences of length 4 (*top*), 5 (*middle*), and 6 (*bottom*), where high error (s.e.m.>20%) history sequences, which correspond to infrequent sequences, have been excluded. Exclusion of these sequences is necessary since, if a sequence is expressed only once it will trivially deterministically evoke a single behavior. Similarly, at low trial numbers, stochasticity is difficult to establish. *Insets:* Confusion matrices for nonparametric policies of each history length for right and left port choice (*left*) and repeat and switch (*right*).



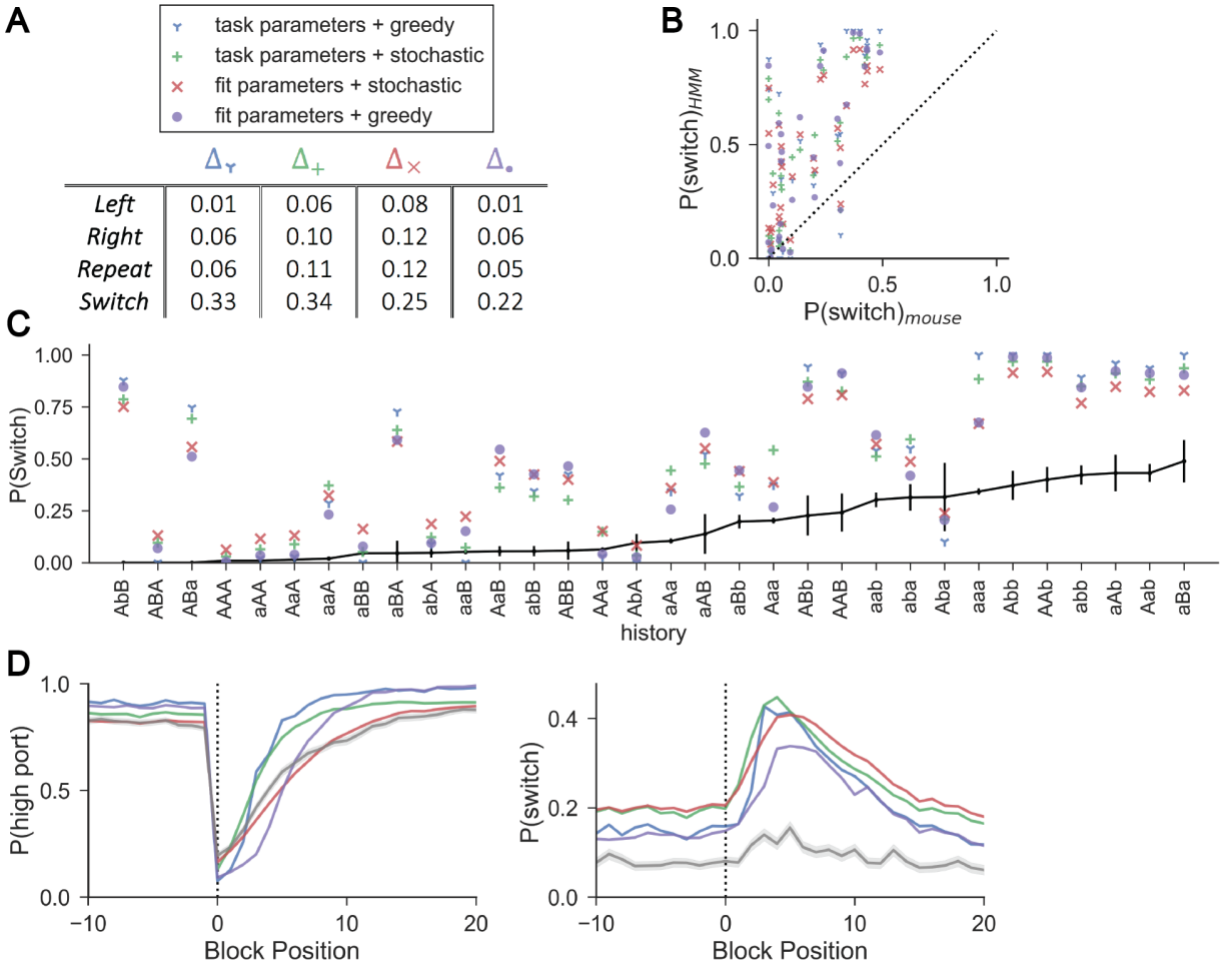
	Δ_{Q1}	Δ_{Q2}	Δ_{Q3}	Δ_{Q4}
Left	0.01	0.01	0.01	0.03
Right	0.01	0.01	0.01	0.01
Repeat	0.00	0.01	0.01	0.00
Switch	0.01	0.00	0.00	0.01

Supplementary Figure 3. Stationarity of behavioral characteristics within sessions. (A) Conditional switch probabilities for each quartile of a session (Q1 to Q4), overlaid on aggregate switch probabilities (black line). Binomial standard error shown for aggregate conditional switch probabilities. (B) Absolute value differences (Δ) for each action from the confusion matrices in Figure 2E (aggregate data) for each within-session quartile. (C) Probability that the mouse chose the high reward probability port (*left*) and switched ports (*right*) as a function of trial number surrounding state transition (block position 0) for each quartile of a session. Dark lines show the mean across trials at the same block position and the shading shows the standard error. (D) Probability density distributions of trial times for each quartile in a session. *Left:* Distribution of time from center port to side port (decision time), *right:* distribution of time from center port to center port (trial initiation to next trial initiation). Probability densities integrate to 1 across the continuous distribution of durations.



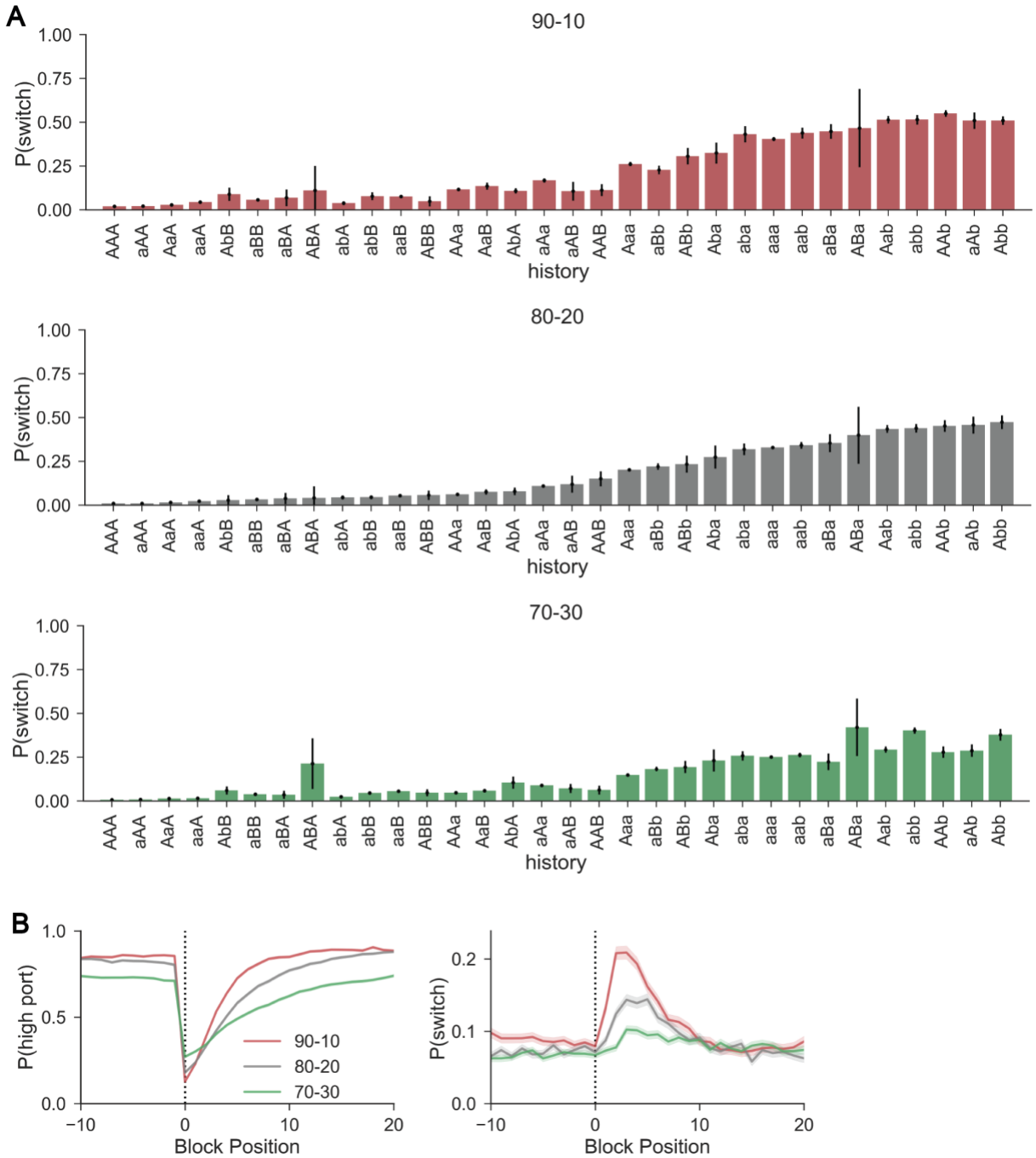
	Δ_{Q1}	Δ_{Q2}	Δ_{Q3}	Δ_{Q4}
Left	0.00	0.01	0.00	0.02
Right	0.01	0.00	0.02	0.01
Repeat	0.01	0.01	0.01	0.00
Switch	0.03	0.00	0.00	0.01

Supplementary Figure 4. Stationarity of behavioral characteristics across sessions. (A) Conditional switch probabilities for session quartiles over the duration of training (i.e., Q1 for early training sessions, Q4 for late training sessions), overlaid on aggregate switch probabilities (black line). Binomial standard error shown for aggregate conditional switch probabilities. (B) Absolute value differences (Δ) for each action from the confusion matrices in Figure 2E (aggregate data) for each across-session quartile. (C) Probability that the mouse chose the high reward probability port (*left*) and switched ports (*right*) as a function of trial number surrounding state transition (block position 0) for each across-session quartile. Dark lines show the mean across trials at the same block position and the shading shows the standard error. (D) Probability density distributions of trial times for each across-session quartile. *Left:* Distribution of time from center port to side port (decision time), *right:* distribution of time from center port to center port (trial initiation to next trial initiation). Probability densities integrate to 1 across the continuous distribution of durations.



	$\Delta_{\curlyvee}$	Δ_{+}	$\Delta_{\times}$	$\Delta_{\bullet}$
Left	0.01	0.06	0.08	0.01
Right	0.06	0.10	0.12	0.06
Repeat	0.06	0.11	0.12	0.05
Switch	0.33	0.34	0.25	0.22

Supplementary Figure 5. Alternative parameterizations of Hidden Markov models fail to capture mouse behavior. (A) Absolute values of the differences between each of the HMM's confusion matrices and nonparametric confusion matrix (Figure 2E). Models, as designated by the labeled indicators: an HMM using the true emission and transition probabilities from the task coupled with a greedy policy, and separately with a stochastic probability matching policy, as well as an HMM using the mouse-fit emission and transition probabilities with each policy. (B) Conditional switch probabilities predicted by each HMM plotted against those observed in mice. Dashed line indicates the unity line. Model indicators correspond to those of (A). (C) Conditional switch probabilities predicted by each HMM, overlaid on the conditional switch probabilities of the mouse (black line). Each data point represents the predicted conditional switch probability by the associated HMM for a given history sequence. Error bars show the binomial standard error for the mouse conditional switch probabilities. Histories are sorted by mouse conditional switch probability. (D) Probabilities of choosing the high reward probability port (*left*) and of switching ports (*right*) as a function of trial number surrounding state transition (block position 0) for each HMM (colors, corresponding to (A)), as well as for the observed mouse behavior (gray). Dark lines show the mean across trials at the same block position, and the shading shows the standard error.



Supplementary Figure 6. Mouse behavior across different reward probability conditions. (A) Conditional switch probabilities for mice performing the task with P(reward|high-choice)=0.9 (*top*), P(reward|high-choice)=0.8 (*middle*, as in Figure 2D), and P(reward|high-choice)=0.7 (*bottom*). For each reward probability set, P(reward|low-choice) = 1-P(reward|high-choice). (B) Summary of probabilities with which mice chose the high reward probability port (*left*) and switched ports (*right*) as a function of trial number surrounding the state transition (block position 0) for each reward probability set. Dark lines show the mean across trials at the same block position and the shading shows the binomial standard error.

Mouse	$P_{highchoice}$ (mean±SD)	P_{reward} (mean±SD)	τ (trials)
A	0.84 ± 0.039	0.70 ± 0.026	5.89
B	0.83 ± 0.050	0.70 ± 0.032	4.45
C	0.81 ± 0.040	0.69 ± 0.024	6.67
D	0.84 ± 0.050	0.70 ± 0.032	6.36
E	0.83 ± 0.044	0.69 ± 0.025	9.17
F	0.82 ± 0.050	0.69 ± 0.030	4.47

Supplementary Table 1. Individual mouse summary statistics for task performance.

Model	Held-out log-likelihood
Nonparametric	-0.180
DDM	-0.180
Sticky HMM	-0.182
Logistic regression	-0.182
HMM, fit	-0.325
HMM, true	-0.359

Supplementary Table 2. Log-likelihoods of held-out data given each model. The nonparametric model uses history length 3, and the logistic regression is the reduced logistic regression with six input features. 'HMM, true' refers to the Thompson sampling HMM with the true task parameters, and 'HMM, fit' refers to that with the mouse-