State-dependent Reward Encoding in Cortical Activity During Dynamic Foraging

Summary. Multiple brain regions are involved in integrating reward to drive action selection, with rich representations of reward history in the striatum, retrosplenial cortex, and frontal areas. A major challenge in dissecting the circuits that govern reward-guided behavior is the existence of multiple strategies for reward maximization. For instance, in dynamic environments, mice can engage in both model-free behavior, where they update action values from trial to trial, and inference-based learning, where they use an internal model to infer the current world state. These two modes are challenging to distinguish, and can even intermix within training sessions, complicating studies of neural mechanisms that rely on session-averaged activity. Here, we tackled these problems by developing a computational approach to characterize dynamic shifts in behavioral strategies. We first simulated the choice sequences of model-free and inference-based agents, and built decoders of their underlying strategy using features of the choice transition around the block switches. We built on this analysis with a new state-space method, block Hidden Markov Model, which infers the hidden state that governs the behavior in each block of trials. Our analysis revealed a diverse mixture of both model-free and inference-based strategies even in expert animals, with an increased reliance on inference-based behavior with training. We used 1-photon widefield imaging to investigate how mesoscopic cortical activity varies with the inferred hidden state. We found that reward encoding is strongly state-dependent: reward is weakly encoded in the disengaged state, transiently encoded in the model-free state, and persistently encoded in inference-based learning. Activity in diverse cortical regions, including the somatosensory, motor, frontal and visual areas, showed different patterns of correlation with reward in each mode. Our results suggest distinct neural mechanisms that underlie different modes of dynamic foraging, and highlight the importance of hidden states in the dissection of reward circuits.

Signatures of model-free and inference-based behavior. We trained head-fixed mice on a dynamic foraging task which alternates between two states, where left or right wheel turns were more likely to be rewarded (Fig. 1a). To determine whether model-free and inference-based behavior can be dissociated based on the animals' choice sequences, we simulated the behavior of Q-learning agents (parameterized by the learning rate and exploration) and inference-based agents (parameterized by its internal model). We measured four features of the choice transition function (Fig. 1b), and identified five distinct regimes in the Q-learning and inference-based spaces (Fig. 1c) which can be reliably decoded based on the choice sequences (Fig. 1d).

Block Hidden-Markov Model reveals shifts in behavioral strategies in single sessions. Although the session-average behavior of mice matched the model-free strategy, we found that this session-average masks the use of multiple strategies within single sessions. We developed a computational method, block Hidden Markov Model (blockHMM), to classify the behavioral strategies in single blocks. This method reveals four hidden states with distinct underlying choice switching dynamics: (1) random behavior, (2) delayed switching, consistent with model-free learning (3) immediate switching, consistent with inference-

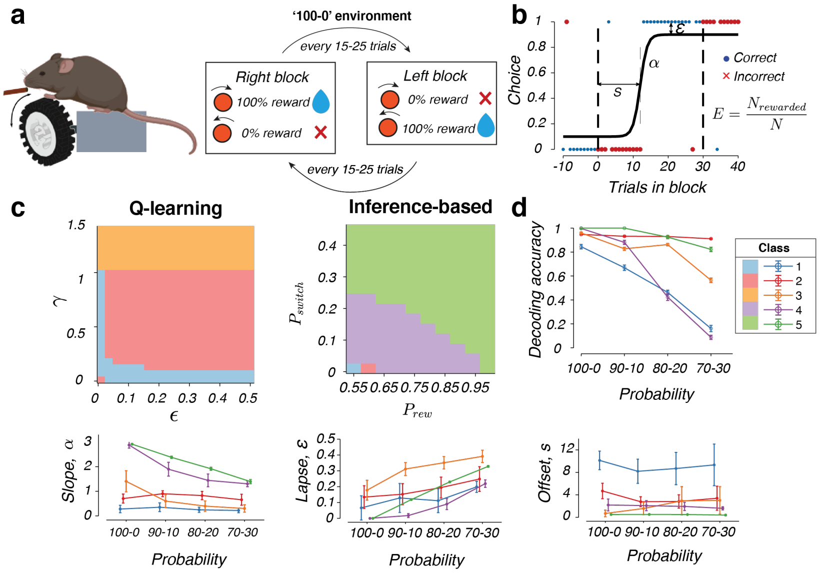
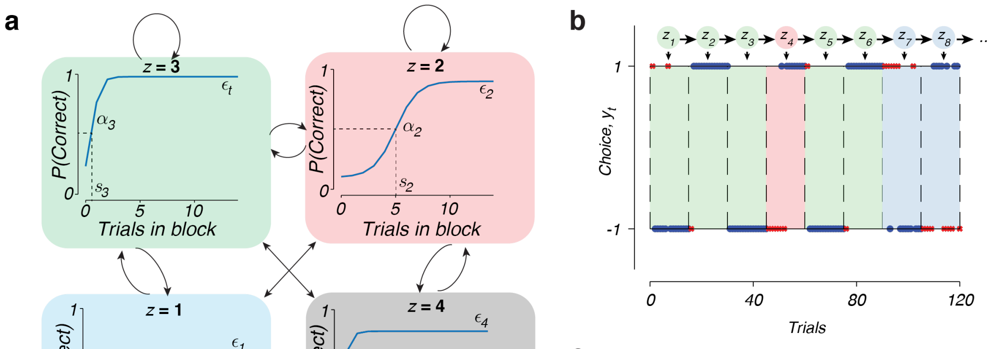


Fig 1. Model-free and inference-based behavior in dynamic foraging. a) Dynamic foraging task (left), and trial structure (right) for head-fixed mice. Mice were trained on 100-0, 90-10, 80-20 and 70-30 environments. **b)** The session-average behavior of rodents and simulated agents might be represented by a sigmoidal transition function with four parameters: switch delay s, switch slope α, exploration (lapse) ε, and overall foraging efficiency, E **c)** Computational simulation of Q-learning and inference-based agents. (Top) Q-learning and inference-based spaces of parameters for the two types of agents. (Bottom) Average behavioral features (mean ± standard error) of agents belonging to each performance regime. **d)** Decoding accuracy of cluster identity using a kNN classifier ($k = 5$).

based strategy, and (4) fast switching but with high lapse rate (Fig. 2). We found that animals employed a mixture of strategies even at the expert stage, and with training, there was a decrease in the random mode and an increase in inference-based, fast-switching mode.



c

BlockHMM equations

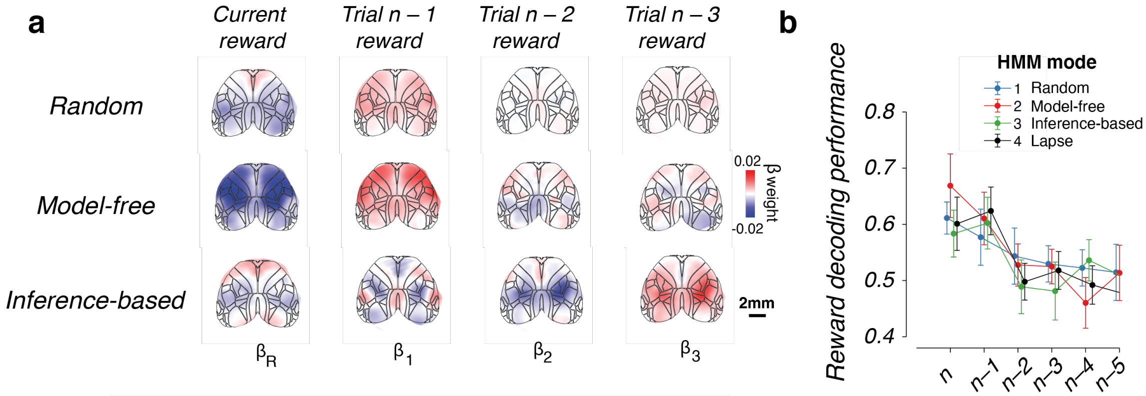
(1) $$\log P(\mathbf{y} \mid z_i = k) = \sum_{t=1}^{T} \log L_k(y_t \mid \alpha_k, s_k, \epsilon_k)$$

(2) $$\log L_k(y_t \mid \alpha_k, s_k, \epsilon_k) = y_t \log \sigma_k(t) + (1 - y_t) \log(1 - \sigma_k(t))$$

(3) $$\sigma_k(t) = \frac{1 - 2\epsilon_k}{1 + \exp\left(\alpha_k(t - s_k)\right))} + \epsilon_k$$

Fig 2. Block Hidden Markov Model. a) Schematic of block Hidden Markov Model (blockHMM). The vector of all choices in a block, $\boldsymbol{y}$, is governed by a hidden state, z_i, whose value transitions from block to block according to transition matrix M (indicated by arrows). Each $z_i = k \epsilon \{1, 2, ..., K\}$ defines a different transition function which is a sigmoidal function with slope α_k, offset s_k, and lapse ε_k. **b)** Example behavior generated by the model in **a**. The circles on top represent the hidden states, z_i, which evolve according to a Markov chain. **c)** The likelihood equations of the generative model. Each hidden state governs the choice sequence of the entire block according to the sigmoidal transitions in **a** (see equations 2 and 3). The log-likelihood of the observed choices in the block is the sum of the log-likelihoods of individual trials (equation 1).

Cortical encoding of reward and reward history is state-dependent. Using the hidden states inferred by blockHMM, we evaluated the cortical representation of reward associated with each behavioral mode. We used 1-photon widefield imaging to record the cortex-wide activity during the task, and examined the relationship between cortical activity and rewards on current and previous trials, split across the random, model-free and inference-based modes. The regression coefficients corresponding to these trial subsets showed a strong dependence on the underlying state. In the random blocks, cortical activity correlates weakly with reward and showed little history encoding (Fig. 3a, top row). In model-free blocks, a large part of the cortex and especially the somatosensory-motor regions were strongly negatively correlated with current reward, but showed short-term history encoding (Fig. 3a, middle row). This short-term reward representation was also seen in reward decoding analysis, which showed a higher decoding of current reward in model-free trials (Fig. 3b). In contrast, inference-based blocks (fast-switching) involved a persistent encoding of reward across past trials (Fig. 3a, bottom row). Inference-based behavior also involved a more positive reward correlation in frontal cortical areas. These analyses support a model of cortical reward integration that is strongly state-dependent: reward encoding in the cortex might be attenuated in the disengaged state, transiently represented in model-free learning for trial-to-trial updates, and integrated and maintained over multiple trials to contribute to model-based inference.



$$Activity(t) = \beta_0 + \beta_R R(t) + \sum_{i=1}^{5} \beta_i R(t-i) + \epsilon$$

Fig 3. Dependence of mesoscopic cortical activation patterns on reward and reward history. a) Results of pixel-based regression of whole-cortex activity against current and previous trial outcomes. Here R(t) = 1 for rewarded trials and 0 for error trials. Regression was done separately for each behavioral mode. Lapse mode (mode 4) was not present for this session. **b)** Decoding of reward and reward history based on the activity of six cortical regions (M1, M2, somatosensory, retrosplenial, PPC and V1). Shown are average performances (mean ± standard error) across $n = 6$ animals (22 sessions).