Neural dynamics in the orbitofrontal cortex reveal cognitive strategies

Shannon S. Schiereck[1], Danilo Trinidad Pérez-Rivera[1], Andrew Mah[1], Margaret L. DeMaegd[1], Royall McMahon Ward[1], David Hocker[1], Cristina Savin[1†], Christine M. Constantinople[1*]

[1] Center for Neural Science, New York University; New York, NY 10003.

† Center for Data Science, New York University; New York, NY 10003.

*Corresponding author. E-mail: constantinople@nyu.edu.

Behavior is sloppy: a multitude of cognitive strategies can produce similar behavioral read-outs. An underutilized approach is to combine multifaceted behavioral analyses with neural recordings to resolve cognitive strategies. Here we show that rats performing a decision-making task exhibit distinct strategies over training, and these cognitive strategies are decipherable from orbitofrontal cortex (OFC) neural dynamics. We trained rats to perform a temporal wagering task with hidden reward states. While naive rats passively adapted to reward statistics, expert rats inferred reward states. Electrophysiological recordings and novel methods for characterizing population dynamics identified latent neural factors that reflected inferred states in expert but not naive rats. In experts, these factors showed abrupt changes following single trials that were informative of state transitions. These dynamics were driven by neurons whose firing rates reflected single trial inferences, and OFC inac-

tivations showed they were causal to behavior. These results reveal the neural signatures of inference.

Introduction

To survive in dynamic environments, animals cannot exclusively rely on learned stimulus-response associations, but must generalize and form inferences about the world; this process is among the most important and interesting cognitive operations that nervous systems perform. The orbitofrontal cortex (OFC) in rodents and primates is implicated in state inference when task contingencies are partially observable[1–6], and when values must be inferred based on high-order associations[7]. How local circuit dynamics in OFC support state inference, however, remains unclear.

For any cognitive computation, including state inference, there are many possible heuristics or alternative strategies that could be used to approximate it[8,9]. A major focus in psychology and neuroscience is to identify the psychological processes that animals (including humans) use to solve cognitive tasks. This is a hard problem, in part because behavioral read-outs in cognitive tasks are often low dimensional (e.g., choice probability, reaction time). Moreover, the space of possible process models is expansive, and many generate qualitatively similar behavior, especially for low dimensional read-outs. Often, behavior on only a small subset of trials is truly diagnostic of different strategies[10,11]. In the limit, e.g., for single-shot inferences or outcome devaluation, only a single trial is used to identify or rule out particular cognitive strategies.

An aspirational goal would be to use rich, multifaceted behavioral read-outs in combination with neural recordings to help constrain the classes of strategies (i.e., process models) that are behaviorally expressed. This approach requires strong behavioral diagnostics of different strategies and neural signatures of cognitive computations that support different model classes. Here, we use multiple, independent lines of evidence from analysis of behavior and large-scale neural recordings to adjudicate between different classes of psychological process models of behavior.

Results

Behavioral evidence for distinct strategies over training.

We developed a temporal wagering task for rats, in which they were offered one of several water rewards on each trial, the volume of which (5, 10, 20, 40, 80μL) was indicated by a tone[12] (Figure 1A). The reward was assigned randomly to one of two ports, indicated by an LED. The rat could wait for an unpredictable delay to obtain the reward, or at any time could terminate the trial by poking in the other port ("opt-out"). Reward delays were drawn from an exponential distribution, and on 15-25 percent of trials, rewards were withheld to force rats to opt-out. How long rats waited before opting out provides a robust analog behavioral readout of their subjective value of the offered water reward[12–15]. Rats were trained in a high-throughput behavioral training facility using computerized, semi-automated procedures to generate statistically powerful datasets across hundreds of animals[12] (N=349 rats).

The task contained latent structure: rats experienced blocks of 40 completed trials (hidden states) in which they were presented with low (5, 10, or 20μL) or high (20, 40, or 80μL) reward volumes[12,14]. These were interleaved with mixed blocks which offered all rewards (Figure 1B). The hidden states differed in their average rewards and therefore in their opportunity costs, or what the rat might miss out on by continuing to wait. According to foraging theories, the opportunity cost is the long-run average reward, or the value of the environment[16]. In accordance with these theories[16,17], well-trained rats adjusted how long they were willing to wait for rewards in each block, and on average waited 10% less time for 20μL in high blocks, when the opportunity cost was high, compared to in low blocks (Figure 1C).

Expert rats' wait time behavior reflected an inferential strategy in which they inferred the

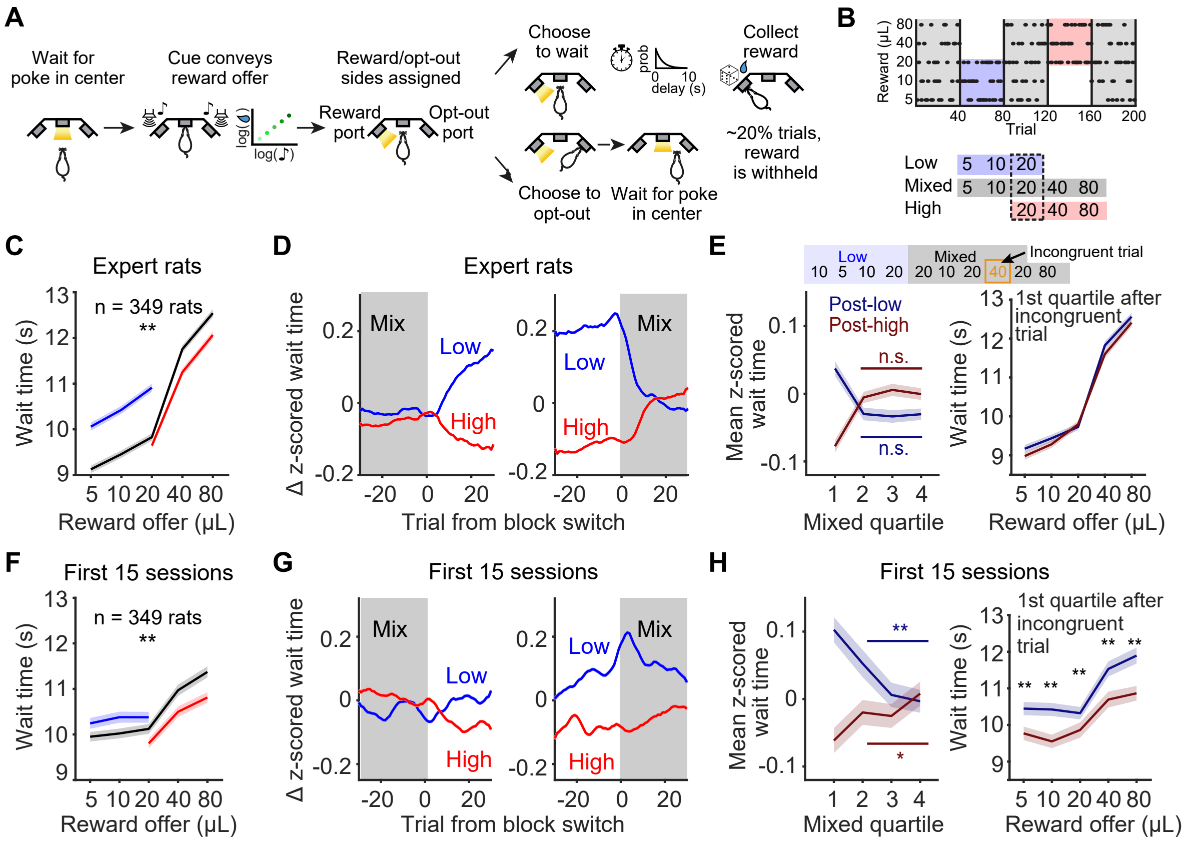

Figure 1: **Behavioral evidence for distinct strategies over training.** **A.** Schematic of behavioral paradigm. **B.** Block structure of task. **C.** Mean wait time on catch trials by reward in each block averaged across expert rats. $p << 0.001$, Wilcoxon signed-rank test comparing wait times for 20μL in high versus low blocks across rats. **D.** Mean (+/-s.e.m.) change in wait time at block transitions from mixed blocks into high or low blocks (left) and high or low blocks into mixed blocks (right), N = 349. Data were smoothed with a causal filter spanning 10 trials. **E.** *left*, Wait times within different quartiles of mixed blocks for expert rats. p-values for effect of quartiles 2-4 on wait times from one-way ANOVA, post-low $p = 0.83$, post-high $p = 0.19$. *right*, Wait times in the first quartile of mixed blocks after the first incongruent trial, which signals a block switch. Curves are conditioned on the previous block type. Bonferroni-corrected p-values for Wilcoxon signed-rank test comparing wait times conditioned on previous block type: 5μL $p = 0.44$, 10μL $p = 0.49$, 20μL $p = 0.16$, 40μL $p = 0.06$, 80μL $p = 0.48$. **F.** Mean wait time by reward in each block in the first 15 sessions of experiencing the blocks. $p = 1.1 \times 10^{-13}$, Wilcoxon signed-rank test comparing wait times for 20μL in high versus low blocks. **G.** Mean (+/-s.e.m.) change in wait time at block transitions from mixed blocks into high or low blocks (left) and high or low blocks into mixed blocks (right), in the first 15 sessions of experiencing blocks, N = 349 rats. Data are plotted as in panel D. **H.** *left*, Wait times within different quartiles of mixed blocks in the first 15 sessions of experiencing the blocks. Data are mean +/- s.e.m. p-values for effect of quartiles 2-4 from one-way ANOVA, post-low

$p = 4 \times 10^{-5}$, post-high $p = 0.02$. *right*, Wait times in the first quartile of mixed blocks after the first incongruent trial, conditioned on the previous block type. Bonferroni-corrected p-values for sign-rank test comparing wait times conditioned on previous block: 5μL $p = 0.002$, 10μL $p = 6 \times 10^{-5}$, 20μL $p = 0.006$, 40μL $p = 5 \times 10^{-5}$, 80μL $p = 1.7 \times 10^{-4}$.

reward block and use a fixed estimate of opportunity cost based on that state inference[12]. This model outperformed alternative process models, and accounted for the dynamics with which rats adjusted their wait times at block transitions (Figure 1D), the insensitivity of their wait times to previous reward offers within a block (Figure S1), and the dependence of their wait times on task parameters such as the catch probability[12]. However, we sought additional behavioral read-outs that might support or falsify the inference hypothesis. We reasoned that an inferential strategy would produce stable wait times in mixed blocks once the animals inferred that the block had changed. To test this, for each rat, we first z-scored the wait times for each reward independently, before pooling over trials with different reward offers. We then computed the mean z-scored wait times in each quartile of mixed blocks that were preceded by low versus high blocks. Consistent with a state inference strategy, rats changed their behavior abruptly, within the first quartile of the mixed block, and then exhibited stable wait times (Figure 1E, *left*). Inferences at transitions into mixed blocks were likely driven by trials offering rewards that were not present in the previous block, which we refer to as incongruent trials (e.g., 40/80μL after a low block, or 5/10μL after a high block). Experts' wait times in the first quartiles of mixed blocks *after* the first incongruent trial were identical regardless of the previous block, consistent with rats inferring a transition into a mixed block following these highly informative trials (Figure 1E, *right*).

Wait times became increasingly sensitive to the hidden states over training, consistent with a strategy that relies on learned task structure[12]. Therefore, we next analyzed the first 15 sessions during which rats were exposed to the blocks, regardless of behavioral performance. Remarkably, even in the first 15 sessions of experiencing the blocks, their wait times showed modest but significant block sensitivity (Figure 1F). However, the behavioral dynamics at block transitions appeared qualitatively different than after extensive training, suggesting a distinct psychological mechanism. Specifically, early in training, while rats showed weak changes in wait times as they transitioned from mixed into high or low blocks, behavioral changes were less apparent when they transitioned from high or low blocks into mixed blocks. Instead, the most striking behavioral feature was an offset or “DC shift” in wait times that persisted into the mixed block, possibly suggesting integration of reward history on longer timescales (Figure 1G). In contrast to expert behavior, early in training, rats’ wait times exhibited prominent within-block dynamics, suggestive of an incremental process of adjusting to the blocks (Figure 1H, *left*). Additionally, wait times in mixed blocks depended on the previous block type, even after the first incongruent trial, further suggesting integration of reward history on long timescales (Figure 1H, *right*). Thus, rats modulate their wait times across latent reward blocks both early and late in training, but analysis of multiple aspects of behavior suggested distinct strategies over training.

Process models of behavior.

We next sought to identify classes of psychological process models that could account for these behavioral observations. The inferential model captured the behavioral dynamics of expert rats at block transitions (Figure 2A,B). The model’s use of fixed, block-specific estimates of opportunity cost reproduced stable wait times in later portions of mixed blocks, and predicted that after the first incongruent trial unambiguously indicated a transition into a mixed block, wait times curves would be identical regardless of the previous block type (Figure 2C). These findings show that the inferential model predicts the behavior of expert rats.

Previous studies have shown that animals can dynamically adjust their subjective value for rewards based on reward statistics via divisive normalization, in which the value of an option is

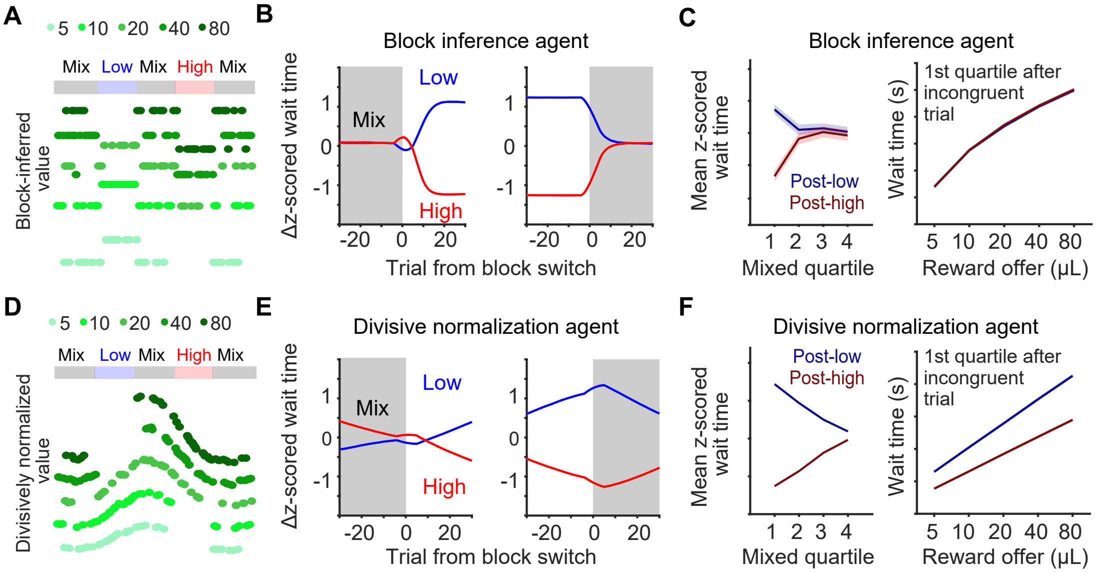

Figure 2: **Psychological process models of behavior.** **A.** Simulated offer values of block inference agent that compares the current reward to a block-specific expectation of average reward, i.e. opportunity cost. **B.** Mean change in wait times from a behavioral model that inferred the most likely block and uses fixed, block-specific values of reward offers to decide how long to wait. **C.** Block inference model predicts that wait times should be fixed within mixed blocks after a block switch has been inferred (left), and that sensitivity to reward offers should not depend on the previous block type (right). **D.** Simulated offer values of divisive normalization agent that divides the value of the current offer by the sum of previous offers in a moving window. **E.** Mean change in wait times for divisive normalization agent. **F.** Divisive normalization model predicts that wait times should change throughout mixed blocks (left), and that value of reward offers in mixed blocks depends on the previous block type. All curves are mean +/- s.e.m.

divided by the sum of previous rewards[14,18,19]. Divisive normalization is a passive process that allows animals to adapt to different stimulus or reward distributions without requiring explicit knowledge of those distributions[19–21]. We simulated the behavior of a divisive normalization agent in our temporal wagering task (Figure 2D). We found that the model captured the key features of behavior early in training, including the modest behavioral changes at transitions into high and low blocks, and the prominent and sustained DC shift in wait times at transitions into mixed blocks (Figure 2E). Divisive normalization predicts incremental changes in wait times throughout the mixed block (Figure 2F), consistent with what was observed early in training (Figure 1H). Finally, within the first quartile of the mixed block, divisive normalization predicts differences in subjective values of rewards (i.e., wait times) depending on the previous block type, even after the first incongruent trial (Figure 2F). For the divisive normalization agent, the incongruent trial is no more or less informative than any other trial, so it fails to produce an abrupt change in the agent's estimate of opportunity cost. These findings show that the divisive normalization model predicts the behavior of rats early in training, when they are naive to the blocks (i.e., "block-naive").

Because divisive normalization is sensitive to the ordering of sequential offers, variability in the sequences of reward offers should influence the degree of block sensitivity in a session[18]. To test this hypothesis, we computed the model's predicted wait time ratio, or the mean predicted wait time for 20μL in a high block divided by a low block, and separated sessions that were in the bottom and top 50th percentiles of wait time ratios. Early in training, rats' block sensitivity was significantly different between these groups of sessions (p=0.003, Wilcoxon signed rank test comparing wait time ratios for sessions predicted to have small or large block effects, N=349). However, in expert rats, block modulation of wait times was not different across these sessions (p=0.34, Wilcoxon signed rank test, N=349). Collectively, these data suggest that early in training, rats adapt their subjective value of rewards to the blocks via a divisive normalization algorithm (or a similar incremental, adaptive process) that integrates over long timescales (tens of trials), and that this process model can explain session-to-session variability in behavioral sensitivity to reward blocks. In contrast, with extended training, rats appear to infer the current block and use fixed, block-specific offer values when deciding how long to wait for rewards.

While in principle, divisive normalization with shorter integration windows could produce faster behavioral changes at block transitions, this model would still predict incrementally changing wait times within a block (Figure S1A,B). Consistent with our previous findings[12], we did not observe such sensitivity to previous rewards in expert animals (Figure S1C). However, this caveat highlights the challenge of definitely ruling out alternative process models of behavior. Therefore, we next sought to test our hypotheses about behavioral strategies using neural recordings.

Latent factors reflect inference in experts.

We performed electrophysiological recordings from the lateral OFC (LO/AI) in block naive and expert rats using chronically-implanted Neuropixels probes (N=42 rats; Figure 3A). These recordings generated large datasets (10,605 single units). Given the scale of these data, we sought to use dimensionality reduction to summarize task-related dynamics. Theoretical models of decision making are often described as low dimensional dynamical systems[22,23], so we focused on low-dimensional neural dynamics, which are also a common statistical feature of neural activity in many contexts[24–29].

While conventional methods for extracting low-dimensional dynamics have focused on the fast (within trial) component of neural activity, a key feature of our task is that determining the value of the reward offer requires integrating over multiple timescales (e.g., evaluating the offer on single trials, inferring the reward block over many trials). To address this limitation, we developed a probabilistic hierarchical linear dynamical systems model (hLDS) that explicitly

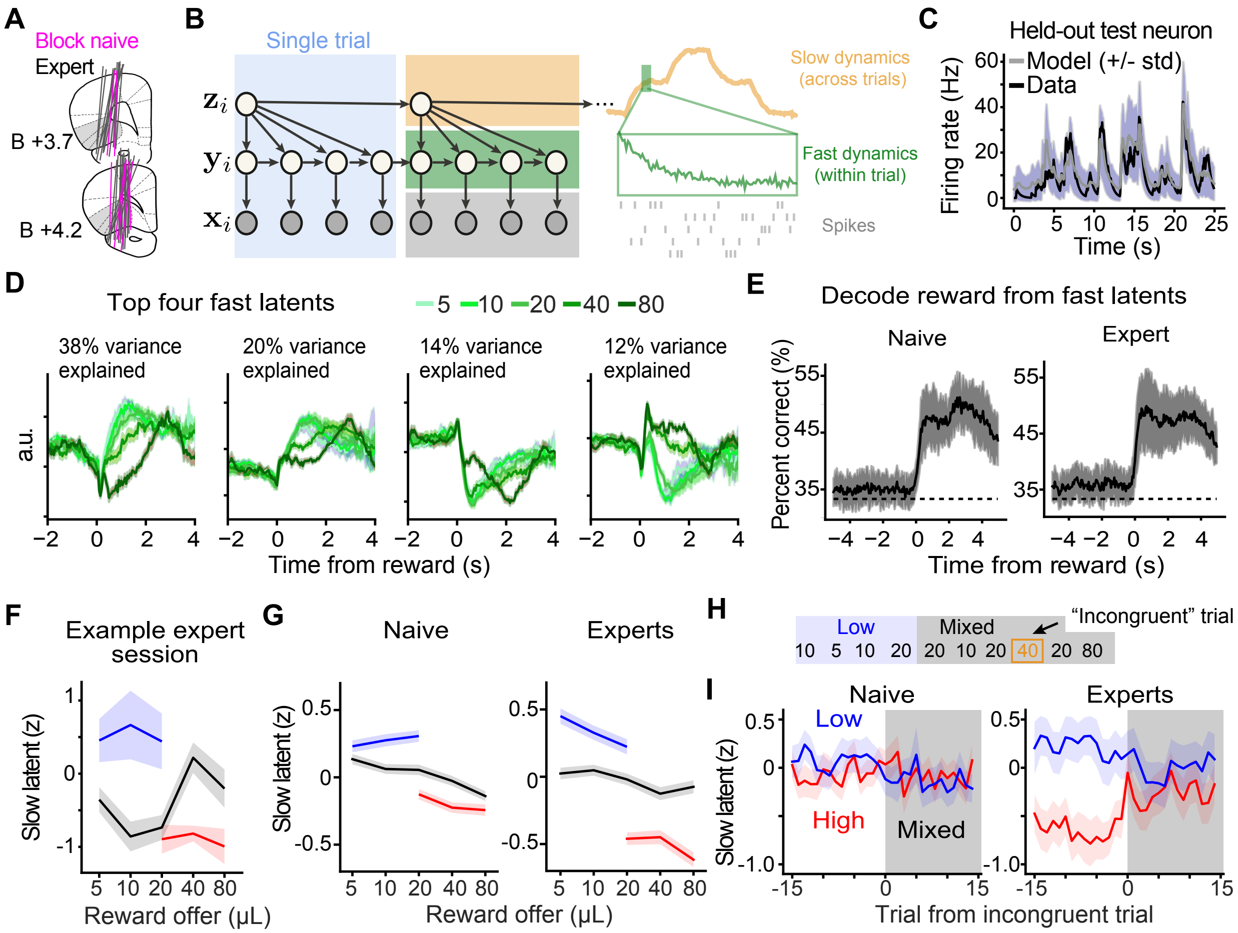

Figure 3: **OFC dynamics reflect inference in expert rats. A.** Location of Neuropixels probe tracks (N = 42 rats). Tracks are shown in a single hemisphere for visualization, but in practice were counterbalanced across hemispheres. **B.** Graphical model of hierarchical linear dynamical systems model (hLDS). For visualization, four fast (within-trial) latents are depicted, but the model was fit using a 1-dimensional z-latent and 10-dimensional y-latents. **C.** Model parameters fit to simultaneously recorded neurons predict the activity of a held-out test neuron. **D.** The four fast latents fit to an example recording session that explain the most variance, aligned to the time of reward for trials with different reward offers. **E.** Performance of a support vector machine decoder, decoding offered reward volume in different time bins around the time of reward. Classifiers decoded whether reward was 5/10, 20, or 40/80, so chance performance was 33% (dashed lines). **F.** Mean slow latent on trials with different reward offers in each block for one example session. **G.** Mean slow latent for block-naive (n=42) and expert (n=58) recordings. Slow latents were z-scored for each session before combining over sessions. **H.** Schematic of incongruent trials, which unambiguously indicate a transition into a mixed block. **I.** Mean slow latent from block-naive and expert recordings, aligned to the first incongruent trial in mixed blocks.

considers multiple interacting timescales. The model assumes a one-dimensional slow latent neural factor (z_k) that operates at the resolution of individual trials, described by a linear gaussian stochastic dynamical system. The fast dynamics within the trial (summarized by 10 dimensional fast latent factors, y_t^k) are assumed to operate in a similar manner. What distinguishes our approach from standard Kalman filtering is that the within trial latent dynamics are themselves dependent on the slower (evolving trial-by-trial) latent process z_k (Figure 3B).

We fit the hLDS model to simultaneously recorded neurons using Expectation-Maximization based parameter estimation (Methods). To validate the model, we showed that it can predict the firing rates of held-out test neurons (Figure 3C), and that it better explains moment-by-moment neural responses than a dimensionality matched standard Kalman filter (Figure S2), suggesting that the hierarchical structure of the dynamics is a key feature of OFC responses during the task. Notably, model-fitting was unsupervised: the model was exclusively fit to the spikes of simultaneously recorded neurons, with no knowledge of the behavioral task. Nonetheless, the fast latents y_t^k captured interpretable features of task-related responses, including the timing of task events and the magnitude of single trial reward offers (Figure 3D). It was possible to decode the reward offer from the fast latent factors, and performance was comparable in both block-naive and expert rats, indicating that knowledge of the blocks was not required for fast-timescale neural dynamics in OFC to reflect rewards (Figure 3E).

The slow latent, z_k, appeared to directly reflect the hidden reward block on individual sessions (Figure 3F) and contained significant mutual information (MI) about the block in the majority of recording sessions in both expert and block naive rats (Figure 3G; expert MI between slow latent and blocks = 0.025, $p << 0.001$; naive MI = 0.01; $p = 0.020$. p-values from non-parametric permutation test, Methods). This suggests that divisive normalization and state inference strategies both result in neural representations of reward blocks in OFC. We reasoned that incongruent trials would be the most diagnostic of whether the z_k latent reflected an incremental, divisive normalization-like process, versus state inference (Figure 3H). We aligned the z_k latent to the first incongruent trial in each mixed block. As described previously, these trials (which do not exist at transitions into high and low blocks) unambiguously reveal that the block has changed. In expert rats, the mean z_k latent showed clear separation before the incongruent trial, and then a sharp convergence to a common value after the first incongruent trial. This was only apparent in recordings from expert rats; block-naive recordings did not reveal a similar abrupt transition (Figure 3I). Therefore, rapid adjustments in latent, population-level neural factors appear to reflect changes in inferred states in expert but not naive animals.

To test the hypothesis that the block sensitivity in naive recordings reflected a different computation, we regressed the z_k latent against previous reward offers in mixed blocks only. While none of the coefficients were significantly different from zero in the expert rats, recordings from block-naive animals had significant regression coefficients for the previous reward offer ($p = 4 \times 10^{-4}$, t-statistic). These data are consistent with incremental, trial-by-trial tracking of reward history in service of an adaptive process like divisive normalization.

Single neuron correlates of state inference.

We next sought to characterize responses to inferred state transitions at the level of individual neurons. We first selected block-sensitive neurons whose firing rates were significantly different in high versus low blocks in the [0 0.5s] window after reward delivery (two-sample t-test, $p < 0.05$). We deemed the block for which they had higher (lower) firing rates the preferred (non-preferred) block for that cell (Figure 4A). Sessions without both transition types (preferred to mixed and non-preferred to mixed) were excluded. We then compared the average firing rates over these neurons for the first congruent or incongruent trial following transitions into mixed blocks. Given that neurons exhibited variable preferences for the different block types, we grouped trials based on whether they indicated a transition away from the neuron's preferred block (non-preferred transition), or away from the neuron's non-preferred block (preferred transition; Figure 4B). In expert rats, neurons exhibited significantly higher firing rates following incongruent trials that indicated preferred transitions, compared to those same trial types in block naive rats ($p = 0.036$, non-parametric permutation test; Figure 4C). The higher firing rates on these individual incongruent trials suggest recognition of a transition away from the non-preferred block. Moreover, there were no differences in firing rates between block-naive and expert rats for congruent mixed block trials (Figure 4C). We interpret the elevated firing rates for congruent trials at non-preferred transitions (compared to preferred transitions) as consistent with rats not yet inferring a transition into a mixed block: because the reward offer is congruent with the previous block, they still believe they are in their preferred (high or low) block. This shows that single trials that are informative of state transitions elicit pronounced increases in the firing rates of individual neurons in the OFC in expert but not block-naive animals. This activity was restricted to the timing of reward delivery, and was not observed at other task events (Figure 4D).

Previous studies in mice have argued that prior beliefs about blocks are represented in all areas of the brain, including early sensory regions[30]. To determine if recognition of incongruent trials was a ubiquitous feature of cortex, we analyzed units that were outside of LO (the Neuropixels probe also traversed through M1 and piriform cortex). We also recorded neurons from the secondary visual area V2. Neurons across all sampled areas seemed to exhibit similar sensitivity to the reward blocks, as classifiers were able to decode the block identity to a comparable degree across brain regions (Figure S3A). However, neither off-target neurons in piriform cortex nor V2 neurons exhibited differential firing rates for incongruent trials (Figure S3B,C). Notably, M1 neurons did exhibit significantly different firing rates for incongruent trials (Figure S3D). We previously found that rats adjust their movement vigor as their beliefs about the reward blocks change, so this could explain neural signatures of inferred state transitions in motor cortex[12,31]. Alternatively, rats could make movements following these highly informative trials, consistent with theories of embodied cognition. Nonetheless, the absence of neural sensitivity to incongruent trials in piriform and visual cortex indicates that this was not a cortex-wide phenomenon (Figure S3).

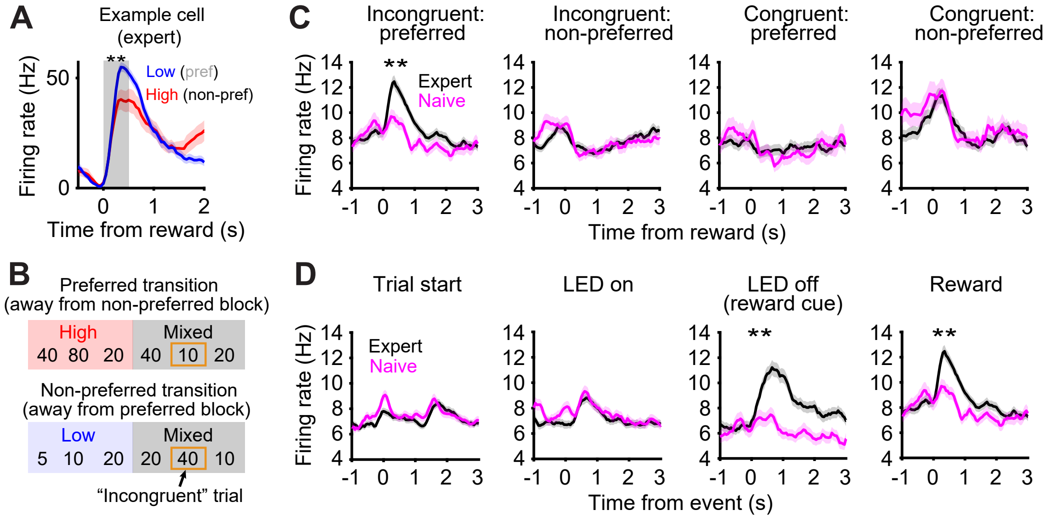

Figure 4: **Single neurons reflect state inference in expert but not naive rats. A.** Example neuron whose firing rate was significantly different in high versus low blocks in the window [0 500ms] after reward. **B.** Explanation of preferred and non-preferred transitions into mixed blocks, for neurons like the example neuron in panel a that prefer low blocks. **C.** Mean (+/- s.e.m) firing rates for neurons with significant block sensitivity (1287/5416) on incongruent and congruent trials after preferred and non-preferred transitions. The expert rat mean is shown in black and the naive rat mean is shown in pink. Asterisks indicate a significant difference in firing rates ($p = 0.036$, Bonferroni-correction, non-parametric permutation test). **D.** Mean firing rates for neurons with significant block sensitivity at trial start (1988/5416), LED on (1323/5416), LED off/reward cue (792/5416), and reward (1287/5416) on incongruent trials after preferred transitions. Expert average is shown in black, naive average is shown in pink. Asterisks indicate a significant difference in firing rates ($p = 0.044$ (LED off), Bonferroni correction, non-parametric permutation test).

We next sought to determine whether this single-cell signature of state inference was broadly distributed across the OFC population, or restricted to specific subpopulations of neurons. To summarize task-related responses at the single neuron level, we used a dimensionality reduction

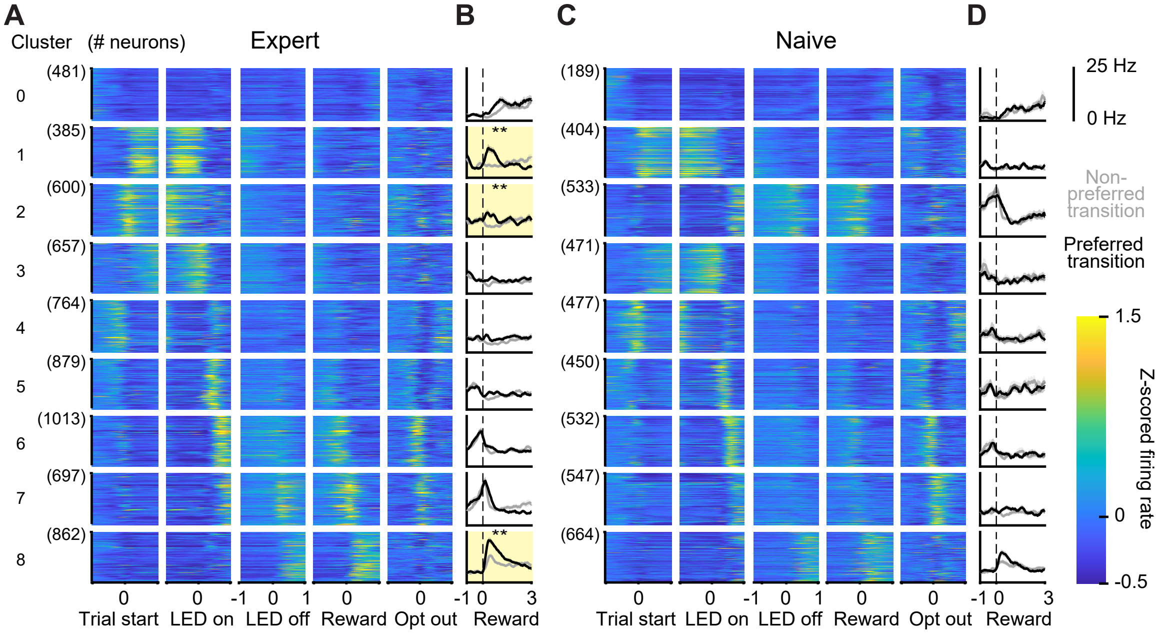

Figure 5: **State inference responses are restricted to subpopulations of neurons A.** Mean event-aligned z-scored firing rates of individual neurons from expert rats, sorted by the TCA component for which they have the maximum loading (see Methods). Parentheticals show number of total neurons in each cluster. **B.** Mean cluster averaged firing rates (raw, not z-scored) at the time of reward delivery for incongruent trials at preferred versus non-preferred block transitions. Yellow boxes and asterisks indicate clusters for which there was a significant difference in firing rates ($p = 0.036$, $p = 0.045$, $p = 0.009$, Bonferroni-correction, non-parametric permutation test). **C.** Same as panel D but for recordings from block-naive rats. **D.** Same as panel E but for block naive rats. No clusters exhibited significantly different firing rates for incongruent trials at preferred versus non-preferred transitions.

method called tensor components analysis (TCA[32]). We constructed a third order data tensor where each row corresponded to the z-scored firing rate of an individual neuron, aligned to different task events; in the z-dimension, we included the neuron's event-aligned activity in each block. Therefore, the data tensor was organized as neurons × time × block, and the model extracted three types of factors: (1) neuron factors, which reflect how much each neuron's activity is described by each component (i.e., loadings); (2) temporal factors, which capture time-varying event-aligned responses, and (3) block factors, which capture modulation of firing rates across blocks. TCA decomposes a third order data tensor into a sum of rank-one components. We selected the number of components based on the number at which adding additional components failed to improve the model fit[32] (Figure S4; see Methods).

We used TCA to perform unsupervised clustering of the neural responses[33]. We clustered neurons by the tensor component for which they had the maximum neuron factor or loading. Neurons that had zero loadings for all components were treated as an additional cluster (cluster 0). In block naive as well as expert rats, the temporal factors for each component captured the mean event-aligned PSTHs for neurons in each cluster (Figure 4D,F). These data are consistent with previous findings that OFC neurons exhibit one of a relatively small subset of temporal response profiles[34,35]. We speculate that these response profiles might act as a temporal basis set for composing dynamics in the OFC. The block factors were generally flat in both groups, indicating that neurons with similar temporal response profiles likely show variable tuning for the reward blocks (Figure S4).

We plotted the cluster-averaged firing rates for incongruent trials following preferred and non-preferred transitions into mixed blocks. Notably, neural encoding of incongruent trials was only apparent in cluster-averaged responses in expert rats, and was restricted to three clusters (1, 2, and 8; Figure 4E). This suggests that sensitivity to incongruent trials at the level of the population firing rate, and also low-dimensional latent neural factors, derives from a subset of neurons that exhibit stereotypical temporal response profiles in the task. Notably, encoding of incongruent trials for these clusters was only apparent at the time of reward and reward cue but not other task events.

OFC inactivations impair state inference.

To determine whether OFC dynamics were causal to state inference, we performed bilateral infusions of the GABA agonist muscimol, targeted to the lateral OFC (LO) in expert rats (Figure 6A). Simultaneous electrophysiological recordings with Neuropixels probes confirmed that muscimol completely silenced neural activity within 1.25mm of the infusion site, indicating that our perturbations silenced LO, agranular insula and ventral OFC, but spared the medial bank of the prefrontal cortex (e.g., PL, IL, CG1; Figure S5A,B).

Inactivating OFC impaired rats' sensitivity to hidden reward states: while in control sessions, animals strongly modulated how long they waited for 20μL in high and low blocks, muscimol reduced this modulation ($p = 0.008$, N = 9 rats; Figure 6B). Moreover, OFC inactivations made rats slower to adjust their wait times following a block transition (Figure 6C). To quantify this effect, we split the high and low blocks into early and late groups of trials (see Methods). In control sessions, sensitivity to hidden reward states was significant early and late in the block, consistent with rapid behavioral adjustments based on state inference. However, in muscimol sessions, the rats were insensitive to hidden states early in the block, although by the end of the block, contextual effects were apparent (Figure 6D). These data show that inactivating OFC did not completely eliminate sensitivity to hidden states, but slowed the dynamics by which rats adjusted their behavior at state transitions.

OFC has been implicated in supporting goal-directed behaviors, as opposed to behaviors that are "model-free" or do not require the use of a world model[3]. Therefore, one possibility is that inactivating OFC caused expert rats to revert to an incremental, trial-by-trial strategy for

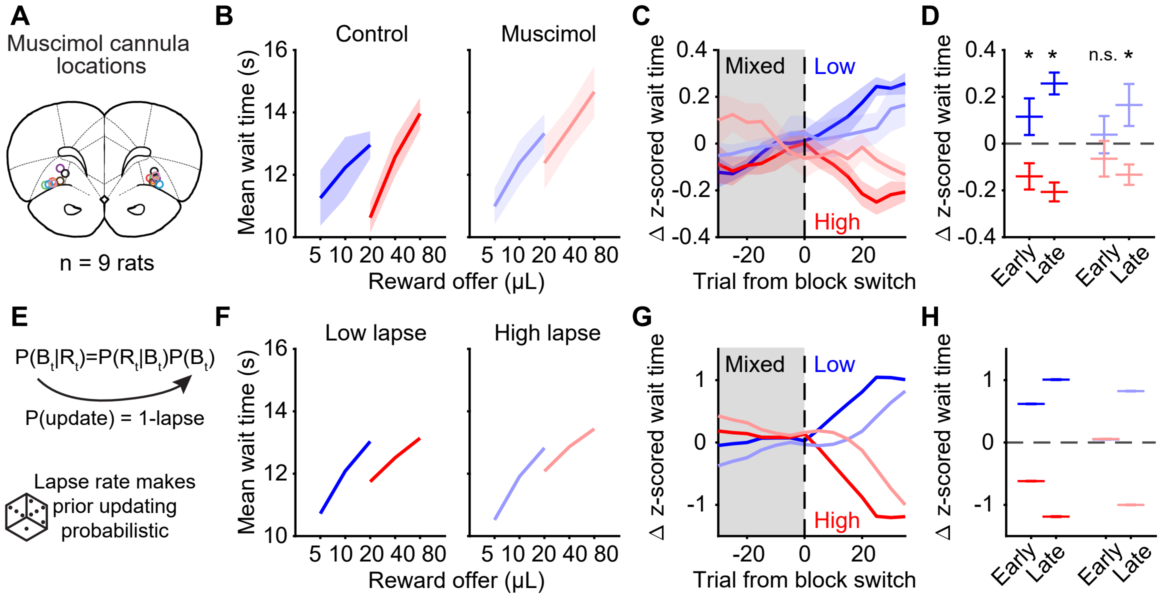

Figure 6: **OFC in expert rats supports belief updating for state inference.** **A.** Location of muscimol guide cannulae in LO (N=9 rats). **B.** Mean wait times in high and low blocks for rats in control or muscimol sessions. Muscimol produces a significant reduction in the wait time ratio (p = 0.008, Wilcoxon sign-rank test). **C.** Mean changes in z-scored wait times as animals transition from mixed into low or high blocks. Dark lines are control sessions and light lines are muscimol sessions. **D.** Mean changes in z-scored wait times early (trials 15-20) and late (trials 35-40) in a block after transitions from a mixed block. Asterisks indicate significant differences between mean z-scored wait times for low and high blocks ($p = 0.039$, control early; $p = 0.004$, control late; $p = 0.027$, muscimol late; Wilcoxon sign-rank test). **E.** The inferential model updates its prior beliefs recursively: the posterior belief on one trial becomes a prior belief on the next trial. We introduced a lapse rate in the model which dictated a probability with which the prior was not updated, and instead remained the same for the next trial. **F.** Increasing the lapse rate (probability of the prior remaining the same) reproduced the reduction in block sensitivity observed with muscimol inactivations. **G,H.** Increasing the lapse rate made the model change its wait time behavior more slowly at block transitions.

estimating the opportunity cost, for instance, via divisive normalization or canonical model-free reinforcement learning[12]. An incremental strategy predicts that wait times in a given block should be sensitive to the magnitude of previous reward offers, potentially for several trials in the past. However, by several measures, rats' wait times in mixed blocks remained insensitive to previous rewards, suggesting that they did not revert to using an incremental adaptive strategy (Figure S5).

We next turned to the inferential model to characterize how OFC inactivations affected behavior (Figure S6A-D). The model uses Bayes' Rule to compute the posterior probability of each block given a reward offer by combining the likelihood, or the probability of encountering the reward in a given block, with the prior belief about the block. The prior over blocks is recursively computed: the posterior on one trial becomes the prior on the next trial[12].

We introduced a lapse rate into the model that dictated the probability with which the posterior became the prior on the next trial (Figure 6E). Increasing this lapse rate increases the probability that the prior on trial t is the prior from t-1 rather than the posterior from trial t-1, in other words, it makes the prior beliefs "sticky." Increasing this lapse rate reproduced the qualitative effects of OFC inactivations, including reduced sensitivity to hidden states and slower behavioral changes at block transitions, while also producing wait times that were largely insensitive to previous rewards within a block (Figure 6F-H). In contrast, reducing the quality of the prior (Figure S6E-G), or making the block-specific opportunity costs more similar (Figure S6H-J), were unable to capture all of the effects of OFC inactivations. These results suggest that OFC supports hidden state inference by updating subjective beliefs based on experience.

We did not perturb OFC in block-naive rats because daily perturbations impair behavioral performance, but intermittent perturbations would allow the rats to learn about the blocks before sufficient inactivation data could be collected. We speculate that such a passive strategy may be distributed and may not causally rely on OFC. Nonetheless, our findings in experts suggest that the dynamical signatures of state inference in these rats were causal to their inferential behavioral strategies.

Discussion

Multiple, independent lines of evidence from behavior and neural recordings indicate that over training, rats transition from passively adapting to reward states via divisive normalization to performing hidden state inference. Multifaceted behavioral analysis was critical for disambiguating between different underlying strategies. Block-naive and expert rats' wait times showed similar sensitivity to the reward blocks. However, examining the dynamics by which wait times changed at block transitions, the dynamics of wait times within mixed blocks, and the sensitivity of mixed block wait times to the previous block type following the first incongruent trial revealed qualitative differences in behavior over training, and led to a more precise characterization of the behavioral deficit following inactivation of OFC. Neural signatures of state inference -abrupt transitions following incongruent trials- were present at the level of single neurons and population-level latent neural factors in expert but not naive rats. Incongruent trials were rare, typically occurring only once or twice in each recording session. We overcame statistical challenges of trial-limited analyses using a brute force approach that included high-throughput training of hundreds of rats, neural recordings of thousands of neurons in dozens of animals (N=42), and dimensionality reduction of neural data.

Divisive normalization is thought to be a canonical computation that supports efficient coding by allowing neurons to adjust the dynamic range of their firing rates to best represent stimulus or reward distributions[21]. While this algorithm is thought to reflect core features of neural circuits like inhibitory motifs and alleviate fundamental constraints of neural coding such as bounded firing rates, expert rats appear to "turn off" divisive normalization in favor of state inference. We speculate that if multiple strategies or neural systems can support behavior through different computations, then each system's relative contribution to the expressed behavior may be determined by a winner-take-all mechanism[22], weighted averaging[10,36], or other arbitration process.

Previous studies in mice have found that task- or behavior-related dynamics are highly distributed and observable in most or all areas of the brain[37–40]. However, just because neural dynamics reflect task-related variables does not mean that those dynamics are causal to behavior[41]. A recent study argued that prior beliefs about blocks (which dictated reward probabilities in a two-alternative forced choice task) were represented brain-wide[30]. That study employed a more permissive definition of prior beliefs that included action repetition, and they found that neural signals reflecting this term were ubiquitous, and observable even in early sensory areas. Similarly, we found that neural activity in all sampled areas including V2, M1, and piriform cortex reflected the hidden reward blocks. However, we refrain from interpreting neural representations of reward blocks *per se* as reflecting computations for state inference, as these representations could reflect many processes including reward history, divisive normalization of value, or even motivation and arousal. By contrast, neural sensitivity to incongruent trials was uniquely observed in the OFC and, to a lesser extent, M1 (among the brain areas we sampled) of expert rats performing our task. This suggests that inferring hidden state transitions likely engages cognitive and neural computations that are preferentially supported by OFC (and that might be reflected in M1). More generally, specific trials or task features that are diagnostic of particular computations may resolve more modular neural representations (i.e., dynamics that are specific to brain areas performing those computations).

A general observation about cortical responses, particularly in the frontal cortex, is that individual neurons respond to diverse combinations of task variables. Studies in the motor system have argued that single neuron heterogeneity derives from variable contributions of individual neurons to population-level latent factors that support the actual computation being performed[42].

While in motor cortex, the computations supported by neural dynamics can be reasonably assumed (e.g., motor preparation and execution), in complex cognitive tasks, there are many quantities and abstract relationships that often must be computed. Theories of mixed selectivity argue that diverse responses at the single neuron level endow downstream circuits with flexibility for decoding different variables depending on changing task demands[43,44]. However, it can be difficult to know which computations are specifically supported by the piece of tissue under study, as well as different downstream recipient circuits. Here, we demonstrated a causal relationship between recorded OFC dynamics and a precise behavioral computation, updating beliefs about hidden reward states, consistent with previous studies[1,3,13,45]. Our unsupervised analysis method revealed population-level neural factors that reflected task computations over multiple timescales, analogous to the motor system but in the context of a cognitive behavioral task. We found that these population-level factors reflected identifiable changes in tuning at the single neuron level, deriving from three functional subpopulations of neurons that reflected single-trial inferences. Neural encoding of incongruent trials was prominent following reward delivery, which may be the task epoch during which belief updating occurs. Collectively, our data identify neural correlates of single trial inferences and show that these dynamics causally update belief distributions over abstract, latent states of the environment.

Methods

Subjects

A total of 349 male and female Long-evans rats between the ages of 6 and 24 months were used for this study (*Rattus norvegicus*). Animal use procedures were approved by the New York University Animal Welfare Committee (UAWC #2021-1120) and carried out in accordance with National Institutes of Health standards.

Rats were pair housed when possible, but were occasionally single housed. Animals were water restricted to motivate them to perform behavioral trials. From Monday to Friday, they obtained water during behavioral training sessions, which were typically 90 minutes per day, and a subsequent ad libitum period of 20 minutes. Following training on Friday until mid-day Sunday, they received ad libitum water. Rats were weighed daily.

Behavioral training

A detailed description of behavioral training has been provided elsewhere[12]. Briefly, rats were trained in a high-throughput behavioral facility in the Constantinople lab using a computerized training protocol. They were trained in custom operant training boxes with three nose ports. Each port contained a visible LED, an infrared LED and infrared photodetector for detecting nose pokes, and the side ports contained lick tubes that delivered water via solenoid vales. There was a speaker mounted above each side port that enabled delivery of stereo sounds. The behavioral task was instantiated as a finite state machine on an Arduino-based behavioral system with a Matlab interface (Bpod State Machine r2, Sanworks), and sounds were delivered using a low-latency analog output module (Analog Output Module 4ch, Sanworks) and stereo amplifier.

Each trial began with the center port being illuminated. Rats initiated the trial by poking their nose in the center point, at which time the light was turn off and an auditory cue would play.

The reward offer on each trial was cued by a tone delivered from both speakers (1, 2, 4, 8, or 16kHz). On each trial, the tone duration was randomly drawn from a uniform distribution from 800ms to 1.2s. Sound pressure was calibrated for each tone (via a gain parameter in software) so that they all matched 70dB in the rig, measured when a microphone (Bruel & Kjaer, Type 2250) was proximal to the center poke. The rat was required to maintain its nose in the center poke for the duration of sound presentation. If it terminated fixation prematurely, that was deemed a violation trial, the rat experienced a white noise sound and time out period, and the same reward offer would be presented on the subsequent trial, to disincentivize premature terminations for small volume offers. Following the fixation period, one of the side LEDs lit up indicating that port would be the reward port. The reward delay on each trial was randomly drawn from an exponential distribution with a mean of 2.5s. When reward was available, the reward port LED turned off, and rats could collect the offered reward by nose poking in that port. On 15-25% of trials, the reward was omitted. The rat could opt out of the trial at any time by poking its nose in the unlit port, after which it could immediately initiate a new trial. In rare instances, on an unrewarded trial, if the rat did not opt-out within 100s, the trial ended ("time-out trial"), and the center LED turned on to indicate a new trial.

We introduced semi-observable, hidden-states in the task by including uncued blocks of trials with different reward offers. High and low blocks, which offered the highest three or lowest three rewards, respectively, were interspersed with mixed blocks, which offered all volumes. There was a hierarchical structure to the blocks, such that high and low blocks alternated between mixed blocks (e.g., mixed-high-mixed-low, or mixed-low-mixed-high). The first block of each session was a mixed block. Blocks transitioned after 40 successfully completed trials. Because rats prematurely broke fixation on a subset of trials, in practice, block durations were variable.

To determine when rats were sufficiently trained to understand the mapping between the auditory cues and water rewards, we evaluated their wait time on catch trials as a function of offered rewards. For each training session, we first removed wait times that were greater than two standard deviations above the mean wait time on catch trials in order to remove potential lapses in attention during the delay period (this threshold was only applied to single sessions to determine whether to include them). Next, we regressed wait time against offered reward and included sessions with significantly positive slopes that immediately preceded at least one other session with a positive slope as well. Once performance surpassed this threshold, it was typically stable across months. Our analysis of expert rat behavior used this criteria to select sessions for analysis. By comparison, to examine behavior early in training, for each expert rat, we analyzed the first 15 training sessions in the final training stage when they first experience the blocks, regardless of behavioral performance.

Training for male and female rats

We collected data from both male and female rats. Male and female rats were trained in identical behavioral rigs with the same shaping procedure (see [12] for detailed description of shaping). To obtain sufficient behavioral trials from female rats who are physically smaller than males, reward offers were slightly reduced while maintaining the logarithmic spacing: [4, 8, 16, 32, 64 μL]. For behavioral analysis, reward volumes were treated as equivalent to the corresponding volume for the male rats (e.g., 16 μL trials for female rats were treated the same as 20 μL trials for male rats). We did not observe any significant differences between male and female rats[12].

Behavioral models

We developed separate behavioral models to describe rats' behavior early and late in training. We adapted a model from [13] which described the wait time, WT, in terms of the value of the environment (i.e., the opportunity cost), the delay distribution, and the catch probability (i.e., the probability of the trial being unrewarded). Given an exponential delay distribution, we defined the predicted wait time as

$$\mathrm{WT} = D\tau \log\left(\frac{C}{1-C} \cdot \frac{R-\kappa\tau}{\kappa\tau}\right).$$

where τ is the time constant of the exponential delay distribution, C is the probability of reward (1-catch probability), R is the reward on that trial, κ is the opportunity cost, and D is a scaling parameter. In the context of optimal foraging theory and the marginal value theorem, which provided the theoretical foundation for this model, each trial is a depleting "patch" whose value decreases as the rat waits[16]. Within a patch, the decision to leave depends on the overall value of the environment, κ, which is stable within trials but can vary across trials and hidden reward states, i.e., blocks.

The inferential model has three discrete value parameters ($\kappa_{\text{low}}, \kappa_{\text{mixed}}, \kappa_{\text{high}}$), each associated with a block. For each trial, the model chooses the κ associated with the most probable block given the rat's reward history. Specifically, for each trial, Bayes' Theorem specifies the following:

$$P(B_t \mid R_t) \propto P(R_t \mid B_t)P(B_t).$$

where B_t is the block on trial t and R_t is the reward on trial t. The likelihood, $P(R_t \mid B_t)$, is the probability of the reward for each block, for example,

$$P(R_t \mid B_t = \text{Low}) = \begin{cases} \frac{1}{3}, & \text{if } R_t = 5, 10, 20\,\mu\text{L} \\ 0, & \text{if } R_t = 40, 80\,\mu\text{L}. \end{cases}$$

To calculate the prior over blocks, $P(B_t)$, we marginalize over the previous block and use the previous estimate of the posterior:

$$P(B_t) = \sum_{B_{t-1}} P(B_t \mid B_{t-1})P(B_{t-1} \mid R_{t-1}). \qquad \text{(Eq. 1)}$$

$P(B_t \mid B_{t-1})$, referred to as the "hazard rate," incorporates knowledge of the task structure, including the block length and block transition probabilities. For example,

$$P(B_t = \text{Low}|B_{t-1}) = \begin{cases} 1 - H_0, & \text{for } B_{t-1} = \text{Low} \\ H_0, & \text{for } B_{t-1} = \text{Mixed} \\ 0, & \text{for } B_{t-1} = \text{High} \end{cases}$$

where $H_0 = 1/40$, to reflect the block length. Including H_0 as an additional free parameter did not improve the performance of the wait time model evaluated on held-out test data in a subset of rats (data not shown), so H_0 was treated as a constant term.

Divisive normalization model

The divisive normalization model divides the value of each offer by the sum of past rewards in some window of trials, following [14]. We modeled the wait times as being directly proportional to this term, by the following equation:

$$WT_t = K \frac{R_t}{1 + \alpha \sum_{k=1}^{N} R_{t-k}}$$

where R_t is the reward offer on trial t, and N dictates the number of previous rewards, and K and α are model parameters. Previous behavioral studies[14] suggested that dynamic valuation in humans was well-captured with an N of 60 previous trials, and this parameter reproduced multiple features of rat behavior early in training. For model simulations, we set $K = 5$ and $\alpha = 0.15$.

When simulating the inferential and divisive normalization models, we treated R_t as $log_2(R_t)$, to be consistent with our previous studies[12], which assumed that rats exhibited compressive utility functions[46]. However, all of our results qualitatively held if we did not log transform the reward offers (data not shown).

Statistical analyses

Exact p-values were reported if greater than 10^{-20}. For p-values smaller than 10^{-20}, we reported $p << 0.001$.

Wait time sensitivity to reward blocks

For all analyses, we removed wait times that were one standard deviation above the pooled-session mean. Without thresholding, the contextual effects are qualitatively similar, but the wait time curves are shifted upwards because of outliers that likely reflect inattention or task disengagement[12]. When assessing whether a rat's wait time differed by blocks, we compared each rat's wait time on catch trials offering 20 μL in high and low blocks using a non-parametric Wilcoxon rank-sum test, given that the wait times are roughly log-normally distributed. We defined each rat's wait time ratio as the average wait time on 20μL catch trials in high blocks/low blocks.

Block transition dynamics

To examine behavioral dynamics around block transitions, for each rat, we first z-scored wait-times for opt-out trials of each volume separately in order to control for reward volume effects. We then computed the difference in z-scored wait times for each volume, relative to the average z-scored wait time for that volume, in each time bin (trial relative to block transition), before averaging the differences over all volumes (Δ z-scored wait time).

For each transition type, we averaged the Δ z-scored wait times and trial initiation times based on their distance from a block transition, including violation trials (e.g., averaged all wait times four trials before a block transition). Finally, for each block transition type, we smoothed the average curve for each rat using a 10-point causal filter, before averaging over rats.

Mixed block quartile analysis

To compute the mean wait times in each quartile of mixed blocks, we first detrended the mean wait time over the course of the session. These effects were modest but in some rats, produced a slight increase in wait times over the session. We regressed mean wait time against trial number pooling over sessions, and subtracted the model-predicted effect of trial number from the wait times of each session. We then z-scored wait-times for opt-out trials of each volume separately in order to control for reward volume effects. We then separated mixed blocks depending on whether they were preceded by a low or high block. We divided each block (including violation trials) into four equally spaced bins of trials. Blocks that were fewer than 40 trials (e.g., if the rat did not complete the block at the end of the training session) were excluded from analysis. We then averaged the z-scored wait times in each quartile/bin for mixed blocks that were preceded by low and high blocks. To determine if there was an effect of mixed quartile on the wait times (i.e., if there were within-block dynamics of wait times), we performed a one-way ANOVA. Because we expected the wait times to change following an inferred state transition in the first quartile, we restricted this analysis to the second through fourth quartiles.

To characterize the mixed block wait times in the first quartile after the first incongruent trial, we first detrended the wait times over the session as described above. We separated mixed blocks depending on whether they were preceded by a low or high block, and divided each block (including violation trials) into four equally spaced bins of trials. We analyzed trials in the first bin/quartile only, and exluded trials preceding and including the first incongruent trial. We then plotted the mean wait times as a function of reward offers for trials in the first mixed block quartile after the first incongruent trial, separately for blocks preceded by low or high blocks. We compared wait times for each reward following a low versus high block using a Wilcoxon signed rank test. To correct for multiple comparisons, we multiplied each p-value by the number of comparisons (five, one for each reward). The p-values reported in the figure legend reflect this Bonferroni correction.

Trial history effects

To assess wait time sensitivity to previous offers (Extended Data Fig. 1b,c), we focused on 20 μL catch trials in mixed blocks only. We z-scored the wait times of these trials separately. Next, we averaged wait times depending on whether the previous offer was greater than or less than 20 μL. For trial initiation times, we used all 20 μL trials in mixed blocks. We averaged z-scored trial initiation times depending on whether the previous offer was greater or less than 20 μL. For both wait time and trial initiation time, we defined the sensitivity to previous offers as the difference between average wait time (trial initiation time) for trials with a previous offer less than 20 μL and trials with a previous offer greater than 20 μL. We compared wait time and trial initiation time sensitivity to previous offers across rats using a paired Wilcoxon signed-rank test.

Neural recordings and analysis

We implanted Neuropixels 1.0 probes in LO (AP +3.7, ML ± 2.5), counterbalanced between left and right hemispheres over rats. Probes were mounted on custom 3-D printed probe mounts[47]. On the day of implantation, probes were lowered so the base of the probe mount sat on the skull (5.5 - 7 mm DV). Animals were allowed to recover for at least five days before recording. Data were acquired using OpenEphys. Spikes were sorted by Kilosort2.0, and manually curated in Phy. Units were further curated using a custom Matlab script. Units with greater than 1% inter-spike intervals less than 1 ms, firing rates less than 1 Hz, or were completely silent for more than 5% of the total recording were excluded. To convert spikes to firing rates, spike counts were binned in 50 ms bins and smoothed using Matlab's smooth.m function.

Before surgery, probes were dipped in the lipophylic dye DiI. Probe tracks were reconstructed from post-mortem histology, and the location of individual recording channels relative to areal boundaries was estimated. Channels that were estimated to be outside of LO or agranular insula (AI) were excluded from further analysis. Cells recorded from channels estimated to be ventral to LO or AI were considered piriform cortex cells. Cells recorded from channels estimated to be dorsal to LO or AI were considered motor cortex cells.

Probes in secondary visual cortex (V2) were implanted at AP -4.7, ML ± 4.0. Channels estimated to be outside of the areal boundaries were excluded from further analysis.

hLDS model

The hierarchical linear dynamical systems (hLDS) model assumes a one-dimensional latent factor $\mathbf{z}_k$ that operates at the resolution of individual trials, described by a linear gaussian stochastic dynamical system:

$$z_{k+1} = Dz_k + u_k, \tag{1}$$

where D is a parameter determining the time scale of the slow dynamics and u_k is independent gaussian white noise, $u_k \sim \mathcal{N}(0, \sigma_u^2)$.

The fast dynamics within the trial, $\mathbf{y}_t^k$ (of dimensionality d) also have linear gaussian dynamics, but driven by the slow component z^k:

$$\mathbf{y}_{t+1}^k = \mathbf{A}y_t^k + Bz^k + w_t, \tag{2}$$

where k and t index the trial, and the time bin within the trial, respectively; the noise w_t is again drawn i.i.d. from a zero mean multivariate normal distribution with isotropic variance, $w_i^k \sim \mathcal{N}(0, \sigma_w^2 I_d)$. The fast dynamics are parametrized by matrix A that determines the recurrent dynamics, vector B that parametrizes the direct influence of the slow latent onto each dimension of the fast dynamics, and noise variance σ_w^2.

Given the fast dynamics, the square-root transformed[25,48] measured spike rates in each time bin are assumed to be generated as a conditionally independent linear gaussian

$$x_t^k \sim \mathcal{N}\left(Cy_t^k, R\right). \tag{3}$$

Parameter matrix C, of size $n \times d$ (where n is the number of simultaneously recorded neurons) determines the degree to which individual neural responses are affected by the low-dimensional population dynamics, with observation noise parametrized by R.

Inference in this model is similar to Kalman filtering/smoothing at each of the layers of the hierarchy. Parameter learning was done by maximum likelihood, via expectation maximization[49]. Smoothing is only used for parameter learning, with filtering used for final latent extraction, to ensure that causal temporal structure is maintained.

The hLDS was fit to sessions for which there were at least 20 simultaneously recorded LO/AI neurons, and the animal completed a full sequence of at least 4 blocks of trials, including at least one low and one high block. The model was fit independently to each session, with a fixed fast latent dimensionality of 10, based on evaluating the dimensionality of eligible sessions by principal components analysis, which consistently suggested diminishing returns in variance explained beyond 10 components (Elbow method).

In order to sort the latent factors by the amount of variance explained, we reparameterized the latent space to produce identical observations by applying a series of linear operations. We leveraged a well-established orthonormalization procedure[25] that uses a singular value-decomposition of the learned observation matrix C, to produce an equivalent parameter set. Under this parameter set, the fast latents are linearly independent, meaning they do not overlap or depend on each other. Then, we sorted these latent variables based on how much variance they explained in the model.

To evaluate model predictions for held-out test neurons, we used the following procedure. The model assumes that the fast latents drive neural activity via the n x d parameter C (where n is the number of neurons and d is the number of fast latents).The held-out neuron's data was included during the fitting procedure, such that an n x d matrix C was learned. For the hold-out test, the row of C corresponding to the held-out neuron was omitted, yielding an estimate of the d-dimensional latent space y using only the n-1 neurons. That is to say, the inference procedure was identical, except using an (n-1) x d = C' and data for all neurons except the held-out neuron. The activity for the left out neuron was then estimated by projecting this inferred latent space y back into the observation space x using the weights from the row that was left out during inference. This procedure was executed on held-out test data that was not used for fitting.

SVM decoder

We constructed a support-vector machine (SVM, using the scikit-learn library in Python) to decode reward volumes from the fast latents extracted from the hLDS. The decoder was trained and tested using trials from all blocks. Fast-latents were discretized into 250 ms time bins. We trained and cross-validated the SVM using 10-fold cross validation. The decoder was trained to decode 5/10, 20, or 40/80, and trials were balanced across groups, so chance performance was 33%.

Mutual information

To determine the relationship between the slow latent process with latent reward blocks across groups of animals, the slow-latent values were grouped across 58 sessions from expert animals and 42 sessions from naive animals. As the sign and magnitude of the slow latent on any given session was arbitrary, these were z-scored across sessions and signed so the mean low-block slow-latent was positive. Mutual information values were computed using a non-binning MI estimator for the case of one discrete data set (reward block) and one continuous data set (z-latent)[50]. We computed significance by shuffling the data labels across n=1000 repetitions, generating a null distribution for our test statistic. Mutual information between the slow latent process and blocks in expert animals (MI = 0.025 ; $p << 0.001$) was greater than the mutual information between the slow latent process and blocks in Naive animals (MI = 0.01; $p = 0.020$). 76% of recording sessions from experts (44/58) contained significant mutual information ($p < 0.01$) about the block when evaluated individually.

Regressing slow latent against reward history

To determine if the slow latent reflected reward history, we first z-scored this variable so that magnitudes were comparable across sessions. We then regressed it against the current reward offer and the previous 10 reward offers (including an offset term), using the built in OLS regression method in the Python statsmodels package. We evaluated the significance of each regression coefficient by the t-statistic. While none of the previous trial coefficients were significant in recordings from expert animals, in naive recordings, the coefficient for the previous reward offer was significantly different from zero ($p = 7 \times 10^{-4}$), suggesting stronger representations of reward history in naive rats.

Single neuron analysis of incongruent trials

To identify neural correlates of inferred state transitions, we first selected neurons that exhibited significantly different firing rates between high and low blocks, in the [0 0.5s] window aligned to the time of reward delivery (two-sample t-test, p<0.05). For these neurons, the block that produced the higher firing rate in that window was deemed the preferred block, and the block that produced the lower firing rate was non-preferred. We then identified transitions from high or low blocks into mixed blocks. For each neuron, we grouped transitions away from the preferred block (non-preferred transitions), and transitions away from the non-preferred block (preferred transitions). Only sessions with both preferred and non-preferred transitions were included, however each transition does not necessarily include both a congruent and incongruent trial. We note that because this was a trial-limited analysis, some neurons only had one transition type, the averages over neurons comprise similar numbers of neurons, but not identical.

To determine if neurons from expert versus naive rats exhibited different firing rates on these trials types we performed a non-parametric permutation test. We generated null distributions on differences in firing rates over groups of rats by shuffling the labels of neurons as belonging to expert or naive animals, recomputing differences in firing rates from randomly drawn groups, and repeating that procedure 1000 times. We computed firing rates in the [0 0.5s] window after reward delivery. We then used this null distribution to calculate a p-value for the observed differences in firing rates between groups of rats: the area under this distribution evaluated at the actual difference of firing rates (between expert and naive rats) was treated as the p-value.

Tensor components analysis

To fit the TCA model, we used software from[32] https://github.com/ahwillia/tensortools and[51] https://github.com/kimjingu/nonnegfac-matlab. We first z-scored each neuron's firing rate, and then fit separate TCA models to all neurons from expert or naive rats. Only neurons from sessions with all three block types were included. Models were fit using non-negative tensor factorization (Canonical Decomposition/PARAFAC). To initially determine the dimensionality, or rank, that should be applied to each model, we iteratively tried different numbers of dimensions, or 'tensor components', and computed the reconstruction error between the model prediction and data. We identified the inflection point, or the point at which adding additional components failed to reduce reconstruction error. Using all of the recorded neurons in each group of animals (expert and naive rats), these error plots suggested that the data were well-captured by a rank 8 model. Adding more than 8 components tended to yield components with flat temporal factors and negligible or zero neuron factors, suggesting that the model was overparameterized.

We grouped neurons based on the component for which they had the highest neuron factor or loading. A subset of neurons in each group had zero loadings for all components. This was because their z-scored firing rates were suppressed throughout the trial, and non-negative TCA failed to capture their task-modulation. We included these neurons as "Cluster 0" in both groups of rats (Fig. 4d-g).

Block decoding

We used multinomial logistic regression to decode the reward block from neuron firing rates [0 0.5s] after reward. We performed 5-fold cross-validation to evaluate decoder performance. Sets of trials used in each training set were balanced across both blocks and volumes. Sessions without all 3 reward blocks or fewer than 5 cells were excluded.

Muscimol infusions

LO was bilaterally inactivated using infusions of muscimol via cannula implanted at AP: +4.0, ML ±2.5 DV -5.0. On muscimol infusion sessions, rats were anesthetized with 2-3% isofluorane in oxygen at a flow rate of 2.5 L/minute and 300-320 nL of muscimol was infused bilaterally through the cannula over a 90s period. Fluid was injected using a Hamilton syringe, and visual confirmation of a drop in the meniscus. Animals were run after a 30-45 minute recovery period. On control sessions, animals were similarly anesthetized but did not receive an infusion of muscimol. Animals were given a two day "wash-out" period to prevent lingering effects of either isofluorane or muscimol. Data for those sessions was not included.

To verify inactivation of neural activity, in an acute experiment, an animal was anesthestized with isoflurane, and a Neuropixels 1.0 probe was lowered at an angle in the same craniotomy as the infusion cannula. Recordings were performed before, during, and up to 30-40 minutes after infusion of 300 nL of muscimol. Based on reconstruction of the probe track from post-mortem histology, we estimated the locations of different recording channels relative to the infusion cannula. We found robust inactivation of neural activity, relative to pre-infusion baselines, up to 1.25mm from the infusion site.

References

1. Vertechi, P. *et al.* Inference-based decisions in a hidden state foraging task: differential contributions of prefrontal cortical areas. *Neuron* **106,** 166–176 (2020).
2. Banerjee, A. *et al.* Value-guided remapping of sensory cortex by lateral orbitofrontal cortex. *Nature* **585,** 245–250 (2020).
3. Wilson, R. C., Takahashi, Y. K., Schoenbaum, G. & Niv, Y. Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81,** 267–279 (2014).
4. Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E. & Behrens, T. E. Frontal cortex and reward-guided learning and decision-making. *Neuron* **70,** 1054–1069 (2011).
5. Stalnaker, T. A. *et al.* Orbitofrontal neurons infer the value and identity of predicted outcomes. *Nature communications* **5,** 3926 (2014).
6. Wallis, J. D. Neuronal mechanisms in prefrontal cortex underlying adaptive choice behavior. *Annals of the New York Academy of Sciences* **1121,** 447–460 (2007).
7. Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science* **338,** 953–956 (2012).
8. Adler, W. T. & Ma, W. J. Comparing Bayesian and non-Bayesian accounts of human confidence reports. *PLoS computational biology* **14,** e1006572 (2018).
9. Bowers, J. S. & Davis, C. J. Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin* **138,** 389–414 (2012).
10. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69,** 1204–1215 (2011).
11. Kool, W., Gershman, S. J. & Cushman, F. A. Cost-benefit arbitration between multiple reinforcement-learning systems. *Psychological science* **28,** 1321–1333 (2017).
12. Mah, A., Schiereck, S. S., Bossio, V. & Constantinople, C. M. Distinct value computations support rapid sequential decisions. *Nature communications* **14,** 7573 (2023).
13. Lak, A. *et al.* Orbitofrontal cortex is required for optimal waiting based on decision confidence. *Neuron* **84** (2014).
14. Khaw, M. W., Glimcher, P. W. & Louie, K. Normalized value coding explains dynamic adaptation in the human valuation process. *Proceedings of the National Academy of Sciences* **114,** 12696–12701 (2017).
15. Steiner, A. P. & Redish, A. D. Behavioral and neurophysiological correlates of regret in rat decision-making on a neuroeconomic task. *Nature neuroscience* **17,** 995–1002 (2014).
16. Charnov, E. L. Optimal foraging, the marginal value theorem. *Theoretical Population Biology* **9,** 129–136 (1976).
17. Stephens, D. W. & Krebs, J. R. in *Foraging theory* (Princeton university press, 2019).

18. Zimmermann, J., Glimcher, P. W. & Louie, K. Multiple timescales of normalized value coding underlie adaptive choice behavior. *Nature communications* **9,** 3206 (2018).

19. Tymula, A. & Glimcher, P. Expected subjective value theory (ESVT): A representation of decision under risk and certainty. *Available at SSRN 2783638* (2021).

20. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nature neuroscience* **4,** 819–825 (2001).

21. Carandini, M. & Heeger, D. J. Normalization as a canonical neural computation. *Nature reviews neuroscience* **13,** 51–62 (2012).

22. Wang, X.-J. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron* **36,** 955–968 (2002).

23. Louie, K., LoFaro, T., Webb, R. & Glimcher, P. W. Dynamic divisive normalization predicts time-varying value coding in decision-related circuits. *Journal of Neuroscience* **34,** 16046–16057 (2014).

24. Paninski, L. & Cunningham, J. P. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. *Current opinion in neurobiology* **50,** 232–241 (2018).

25. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience* **17,** 1500–1509 (2014).

26. Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods* **15,** 805–815 (2018).

27. Zhao, Y. & Park, I. M. Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation* **29,** 1293–1316 (2017).

28. Yu, B. M. *et al.* Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *Advances in neural information processing systems* **21** (2008).

29. Wu, A., Roy, N. A., Keeley, S. & Pillow, J. W. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. *Advances in neural information processing systems* **30** (2017).

30. Findling, C. *et al.* Brain-wide representations of prior information in mouse decision-making. *BioRxiv,* 2023–07 (2023).

31. Mah, A., Golden, C. E. & Constantinople, C. M. Dopamine transients encode reward prediction errors independent of learning rates. *Cell Reports* **43** (2024).

32. Williams, A. H. *et al.* Unsupervised discovery of demixed, low-dimensional neural dynamics across multiple timescales through tensor component analysis. *Neuron* **98,** 1099–1115 (2018).

33. McGuire, K. L. *et al.* Visual association cortex links cues with conjunctions of reward and locomotor contexts. *Current Biology* **32,** 1563–1576 (2022).

34. Hocker, D. L., Brody, C. D., Savin, C. & Constantinople, C. M. Subpopulations of neurons in lOFC encode previous and current rewards at time of choice. *Elife* **10,** e70129 (2021).

35. Hirokawa, J., Vaughan, A., Masset, P., Ott, T. & Kepecs, A. Frontal cortex neuron types categorically encode single decision variables. *Nature* **576,** 446–451 (2019).

36. Miranda, B., Malalasekera, W. N., Behrens, T. E., Dayan, P. & Kennerley, S. W. Combined model-free and model-sensitive reinforcement learning in non-human primates. *PLoS computational biology* **16,** e1007944 (2020).

37. Steinmetz, N. A., Zatka-Haas, P., Carandini, M. & Harris, K. D. Distributed coding of choice, action and engagement across the mouse brain. *Nature* **576,** 266–273 (2019).

38. Chen, S. *et al.* Brain-wide neural activity underlying memory-guided movement. *Cell* **187,** 676–691 (2024).

39. Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* **364,** eaav3932 (2019).

40. Musall, S., Kaufman, M. T., Juavinett, A. L., Gluf, S. & Churchland, A. K. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience* **22,** 1677–1686 (2019).

41. Pinto, L. *et al.* Task-dependent changes in the large-scale dynamics and necessity of cortical regions. *Neuron* **104,** 810–824 (2019).

42. Churchland, M. M. & Shenoy, K. V. Preparatory activity and the expansive null-space. *Nature Reviews Neuroscience* **25,** 213–236 (2024).

43. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology* **37,** 66–74 (2016).

44. Tye, K. M. *et al.* Mixed selectivity: Cellular computations for complexity. *Neuron* (2024).

45. Masset, P., Ott, T., Lak, A., Hirokawa, J. & Kepecs, A. Behavior-and modality-general representation of confidence in orbitofrontal cortex. *Cell* **182,** 112–126 (2020).

46. Constantinople, C. M., Piet, A. T. & Brody, C. D. An analysis of decision under risk in rats. *Current Biology* **29,** 2066–2074 (2019).

47. Luo, T. Z. *et al.* An approach for long-term, multi-probe Neuropixels recordings in unrestrained rats. *Elife* **9,** e59716 (2020).

48. Kihlberg, J. K., Herson, J. H. & Schotz, W. E. Square Root Transformation Revisited. *Journal of the Royal Statistical Society Series C: Applied Statistics* **21,** 76–81. ISSN: 0035-9254. eprint: `https://academic.oup.com/jrsssc/article-pdf/21/1/76/48612658/jrsssc\_21\_1\_76.pdf`. `https://doi.org/10.2307/2346609` (Mar. 1972).

49. Moon, T. K. The expectation-maximization algorithm. *IEEE Signal processing magazine* **13,** 47–60 (1996).

50. Ross, B. C. Mutual information between discrete and continuous data sets. *PloS one* **9,** e87357 (2014).

51. Kim, J. & Park, H. in *High-performance scientific computing: Algorithms and applications* 311–326 (Springer, 2012).

Acknowledgments

We thank Kenway Louie, Tony Movshon, Alex Williams, and members of the Constantinople lab for feedback on the manuscript and helpful discussions.

Funding: This work was supported by a K99/R00 Pathway to Independence Award (R00MH-111926), an Alfred P. Sloan Fellowship, a Klingenstein-Simons Fellowship in Neuroscience, an NIH Director's New Innovator Award (DP2MH126376), an NSF CAREER Award, and a McKnight Scholars Award to C.M.C., R01MH125571 to C.S. and C.M.C., and K01MH132043 to D.H. S.S. and A.M. were supported by 5T32MH019524. A.M. was supported by 5T90DA043219 and F31MH130121. D.P. was supported by a DOE Computational Sciences Graduate Fellowship.

Author Contributions

S.S.S. collected electrophysiology data, with assistance from M.L.D. and R.M.W. S.S.S. performed muscimol experiments, and analyzed electrophysiology and behavioral data. D.P. developed the hLDS, under the supervision of D.H. and C.S. A.M. contributed to behavioral modeling. S.S.S, D.P. and C.M.C. prepared the figures. C.M.C. and S.S.S wrote the manuscript. C.M.C. and C.S. supervised the project.

Data Availability

The data generated in this study will be deposited in a Zenodo database upon publication.

Code Availability

Code used to analyze all data and generate figures will be available at `https://github.com/constantinoplelab/published/tree/main` upon publication.

Supplementary materials

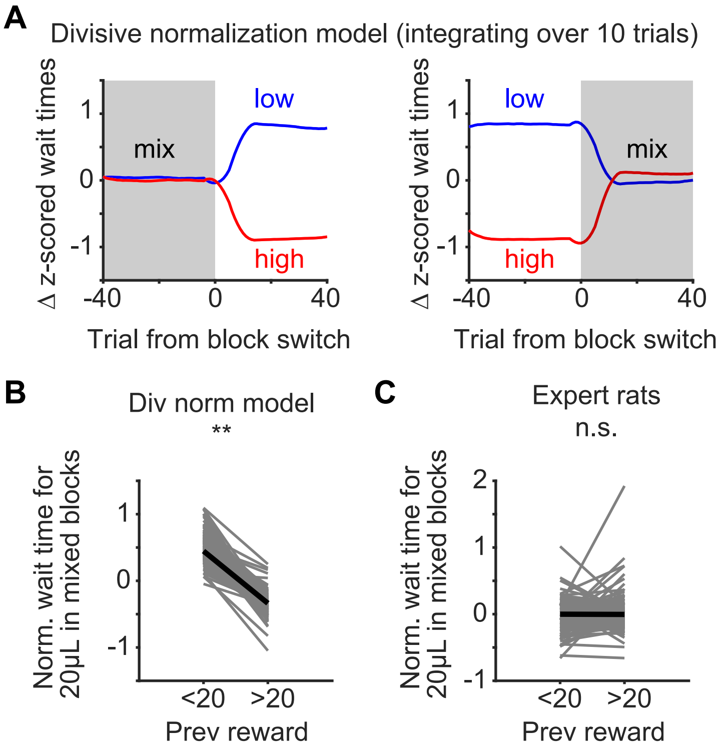

Figure S1 : **Divisive normalization agent with shorter integration windows, related to Figure 2 A.** We simulated the behavior of a divisive normalization agent that integrated over 10 trials (as opposed to 60, which was used throughout the rest of the manuscript). The model was simulated for the trial sequences of each rat, and then predicted wait times were averaged over simulations (i.e., n=349 simulated agents). Data are mean +/- s.e.m. **B.** The divisive normalization model predicts that wait times for the same reward (20μL) in a mixed block should vary depending on whether the previous reward was greater than or less than 20μL. $p << 0.001$, Wilcoxon sign-rank test comparing normalized wait times for 20μL in mixed blocks conditioned on previous reward volume. **C.** Expert rats' wait times for 20μL in mixed blocks conditioned on previous reward volume. $p = 0.06$, Wilcoxon sign-rank test.

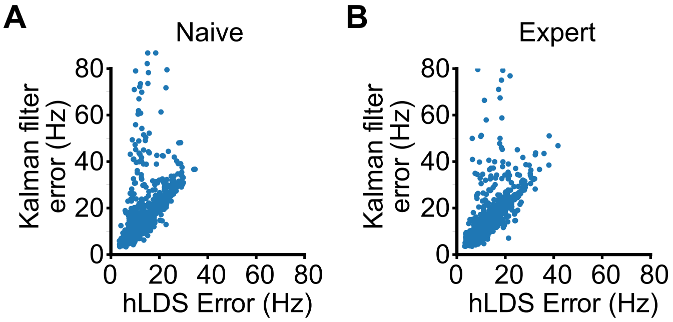

Figure S2 : **Hierarchical LDS outperforms dimensionality-matched Kalman filter, related to Figure 3 A.** Given that the hLDS was fit with an 11 dimensional latent space (10 fast latents, 1 slow latent), an 11-dimensional standard Kalman Filter was also fit to each session. The reconstruction error for left-out neurons on held-out test data was compared across models in recordings from block-naive rats, and favored the hLDS. $p << 0.001$, Wilcoxon sign-rank test. **B.** Same as panel a, but for recordings from expert rats. $p << 0.001$, Wilcoxon sign-rank test.

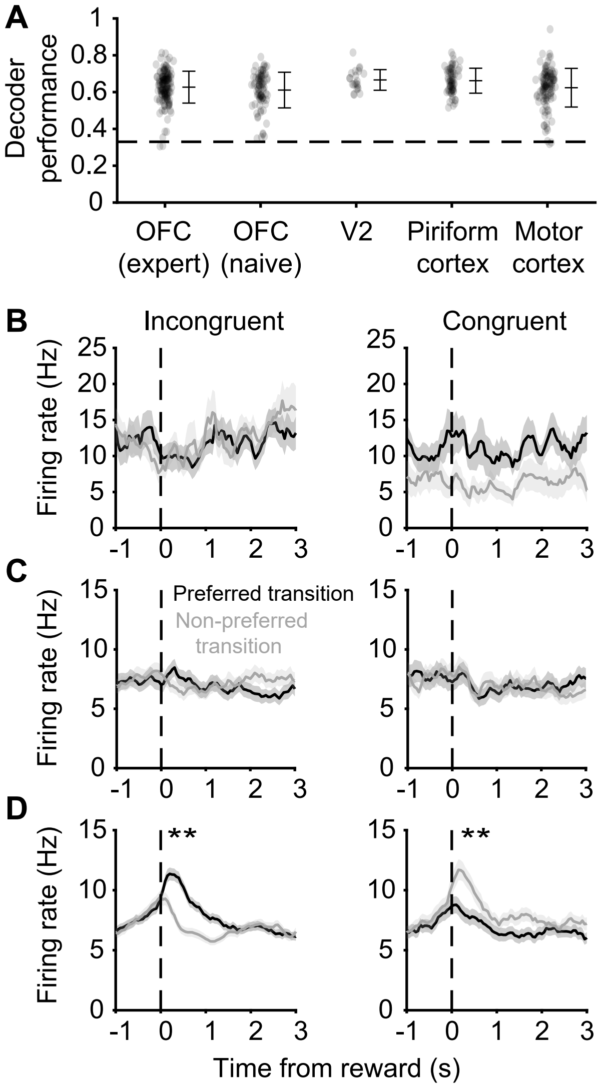

Figure S3 : **Responses to incongruent trials are not ubiquitous, related to Figure 4 A.** Block decoding performance averaged over simultaneously recording neurons in a session. The current reward block was decoded above chance (0.33) in all regions. Each dot represents a recording session. Error bars are mean +/- standard deviation. **B.** Mean firing rates for V2 neurons with significant block sensitivity (n = 38/266) at incongruent and congruent trials signaling preferred transitions (black) and non-preferred transitions (gray). $p = 0.74$ (incongruent), $p = 0.08$ (congruent), Bonferroni-correction, non-parametric permutation test. **C.** Same as panel A but for neurons in piriform cortex with significant block sensitivity (304/1625). $p = 0.54$ (incongruent), $p = 0.65$ (congruent), Bonferroni-correction, non-parametric permutation test. **D.** Same as panel B but for neurons in motor cortex with significant block sensitivity (852/3031). $p << 0.001$ (incongruent), $p = 0.01$ (congruent), Bonferroni-correction, non-parametric permutation test.

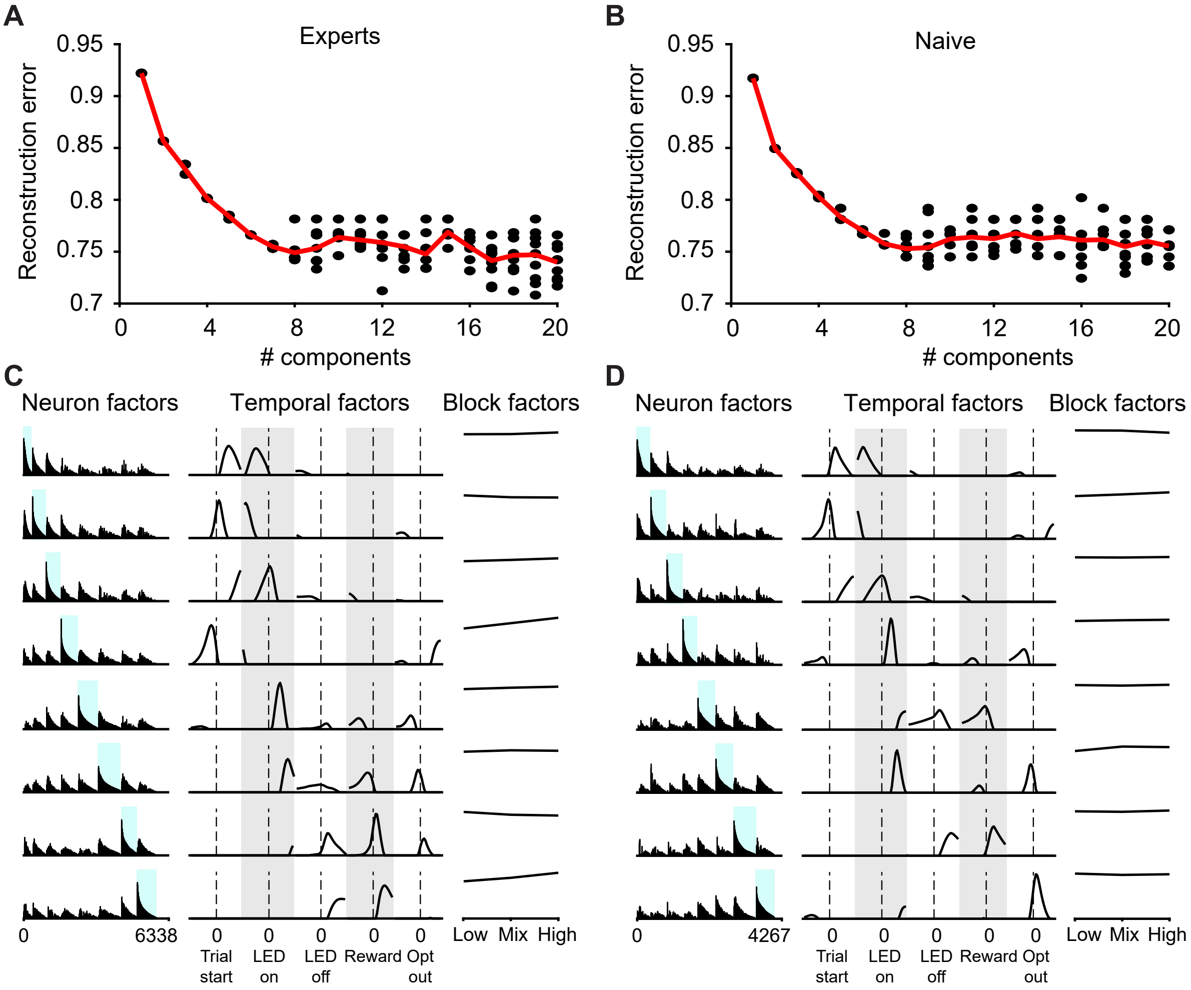

Figure S4 : **TCA reveals 8 clusters of neurons with distinct event-aligned responses, related to Figure 5 A.** To determine the dimensionality, or rank, that should be applied to the neurophysiology data, we iteratively tried different numbers of dimensions, or 'tensor components', and computed the model reconstruction error. In expert rats, more than 8 components failed to improve model performance (elbow method). **B.** Same as panel A but for naive rats. **C.** Neuron factors, temporal factors, and block factors for the TCA model of rank 8 for expert rats. Neuron factors correspond to the weight for each cell. Temporal factors correspond to the average event-aligned responses. Block factors correspond to the magnitude of the response in each block. Components are ordered by the center of mass for the temporal factors. **D.** Same as panel C but for naive rats.

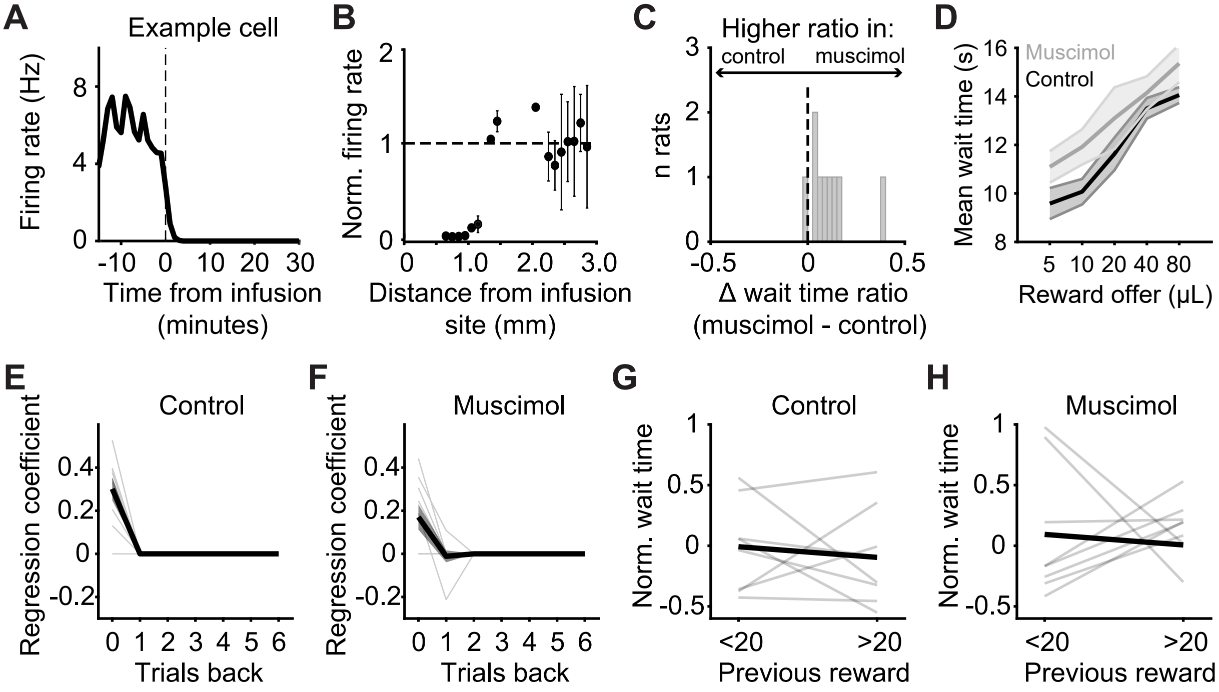

Figure S5 : **Muscimol inactivation of lateral OFC, related to Figure 6 A.** Example neuron recorded within the infusion radius. The neuron is completely silenced within minutes after the infusion. **B.** Average firing rate for neurons in 0.1 mm bins after muscimol infusion. Post-infusion firing rates were normalized to pre-infusion firing rates. Error bars are standard deviation. **C.** Change in wait time ratio between control and muscimol sessions. Bars to the right of 0 indicate a higher wait time ratio (closer to 1) in muscimol sessions compared to control sessions. **D.** Mean wait times across rats for control (black) and muscimol (gray) sessions in mixed blocks. Slopes were not significantly different between groups ($p = 0.164$, Wilcoxon sign-rank test.) **E,F.** Regression coefficients for wait time in control and muscimol sessions. Wait times were regressed against reward offers on the previous trial. **G,H.** Wait time on 20 μL catch trials in mixed blocks conditioned on previous reward offer for control sessions ($p = 0.742$, Wilcoxon sign-rank test) and muscimol sessions ($p = 0.742$, Wilcoxon sign-rank test).

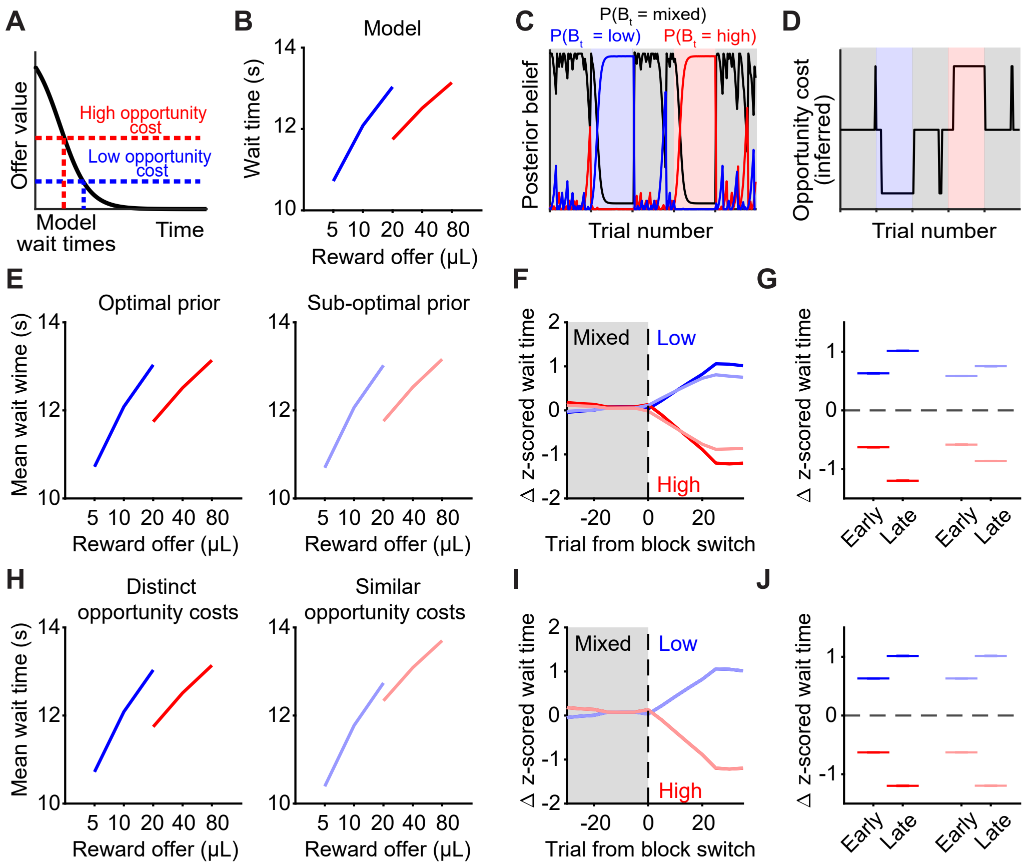

Figure S6 : **Other models are unable to capture OFC inactivation effects, related to Figure 6 A.** Inferential model schematic. **B.** Example model-predicted wait times for low (blue) and high (red) blocks. **C.** Example model-computed posterior beliefs for each block. The model computes a probability for each reward block on each trial. The model predicted block is the one with the highest posterior probability. **D.** Example model-inferred opportunity cost selected based on the maximum posterior belief in C. The model selects from three distinct opportunity costs, one for each reward block. Offer values on each trial are compared to the inferred opportunity cost. **E.** We tested whether inactivation of lateral OFC impairs the quality of the prior by simulating wait times with the inferential model using an optimal prior and a sub-optimal prior. A sub-optimal prior does not reduce the wait time ratio. **F,G.** Simulated mean changes in z-scored wait times early (trials 15-20) and late (trials 35-40) in a block after transitions from a mixed block. Dark colors indicate an optimal prior, light colors indicate a sub-optimal prior. A sub-optimal prior does not change the transition dynamics. Asterisks indicate significant differences between mean z-scored wait times for low and high blocks ($p << 0.001$ for all comparisons, Wilcoxon sign-rank test). **H.** We tested whether inactivation of lateral OFC impairs the ability to distinguish between 3 unique reward blocks with distinct opportunity costs. We simulated wait times using the inferential model with a distinct opportunity cost associated with each block or a similar opportunity cost associated with each block. An agent

with similar opportunity costs associated with each block reduces the wait time ratio. **I,J.** Simulated mean changes in z-scored wait times early (trials 15-20) and late (trials 35-40) in a block after transitions from a mixed block. Dark colors indicate distinct opportunity costs, light colors indicate similar opportunity costs. Using similar opportunity costs for each block does not change the transition dynamics. Asterisks indicate significant differences between mean z-scored wait times for low and high blocks ($p << 0.001$ for all comparisons, Wilcoxon sign-rank test).