Competitive integration of time and reward explains value-sensitive foraging decisions and frontal cortex ramping dynamics

Michael Bukwich[1,2,9]*, Malcolm G. Campbell[1,2]*, David Zoltowski[3,4], Lyle Kingsbury[1,2], Momchil S. Tomov[1,2,10], Joshua Stern[1,2], HyungGoo R. Kim[1,2,7,8], Jan Drugowitsch[6], Scott W. Linderman[3,4], Naoshige Uchida[1,2]

1. Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, 02138
2. Center for Brain Science, Harvard University, Cambridge, MA, 02138
3. Department of Statistics, Stanford University, Stanford, CA, 94305
4. Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305
5. Department of Psychology, Harvard University, Cambridge, MA, 02138
6. Department of Neurobiology, Harvard Medical School, Boston, MA, 02115
7. Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon 16419, Republic of Korea
8. Department of Biomedical Engineering, Sungkyunkwan University, Suwon 16419, Republic of Korea
9. Current address: Sainsbury Wellcome Centre, University College London, London, W1T 4JG, UK
10. Current address: Motional AD LLC, Boston, MA 02210

* These authors contributed equally

Correspondence: uchida@mcb.harvard.edu

Abstract

The ability to make advantageous decisions is critical for animals to ensure their survival. Patch foraging is a natural decision-making process in which animals decide when to leave a patch of depleting resources to search for a new one. To study the algorithmic and neural basis of patch foraging behavior in a controlled laboratory setting, we developed a virtual foraging task for head-fixed mice. Mouse behavior could be explained by ramp-to-threshold models integrating time and rewards antagonistically. Accurate behavioral modeling required inclusion of a slowly varying "patience" variable, which modulated sensitivity to time. To investigate the neural basis of this decision-making process, we performed dense electrophysiological recordings with Neuropixels probes broadly throughout frontal cortex and underlying subcortical areas. We found that decision variables from the reward integrator model were represented in neural activity, most robustly in frontal cortical areas. Regression modeling followed by unsupervised clustering identified a subset of neurons with ramping activity. These neurons' firing rates ramped up gradually in single trials over long time scales (up to tens of seconds), were inhibited by rewards, and were better described as being generated by a continuous ramp rather than a discrete stepping process. Together, these results identify reward integration via a continuous ramping process in frontal cortex as a likely candidate for the mechanism by which the mammalian brain solves patch foraging problems.

Introduction

One of the most fundamental computations animals must regularly make to ensure their survival is how to effectively allocate time in the pursuit of resources. Patch foraging—the process of deciding when to leave a location with depleting resources to search for a new one—is an ethologically relevant decision-making problem faced by many species across the animal kingdom, from worms to primates [1-6]. A classic result in behavioral ecology is that foraging animals should leave a patch when the marginal rate of resource intake equals the average rate in the environment—a result known as the Marginal Value Theorem (MVT) [7]. Theoretical work has demonstrated that integrate-to-threshold models, classically used to model evidence accumulation in sensory decision making, can produce optimal value-based decisions [8], and can be used to solve the patch foraging problem [9]. These integrator models can achieve MVT-optimal behavior and, through parameter tuning, can also implement non-MVT strategies appropriate to different distributions of resources in the environment. The idea that integrator models can be adapted to solve foraging problems is attractive, because they offer a potential mechanism through which to link foraging behavior to its neural basis.

Rodents are natural foragers and are highly amenable to a wide range of molecular/genetic tools for dissection of neural circuit function [10]. Therefore, foraging in rodents presents an ideal opportunity to dissect neural mechanisms of an ethologically relevant cognitive computation with a well-grounded theoretical basis. Recent work has identified multiple areas of rodent frontal cortex [11], including anterior cingulate cortex (ACC) [12], orbitofrontal cortex (OFC) [13], and secondary motor cortex (M2) [14] as being involved in foraging decisions. We therefore set out to record neural activity broadly in the frontal cortex while mice performed a well-controlled laboratory foraging task, aiming to identify the behavioral algorithms, neural activity patterns and brain areas that support this natural decision process.

Here, we present a novel head-fixed virtual reality patch foraging task for mice. By comparing various models of patch-leaving decisions, we show that while mouse behavior did not conform precisely to the MVT, integration models captured mouse behavior with quantitative precision. Critical to accurate behavioral modeling was a slowly varying "patience" variable, estimated from surrounding trials, which modulated sensitivity to time. Decision variables derived from integration models could be predicted from neural activity in the frontal cortex, and vice versa, more accurately than in subcortical areas. These results suggest that mice solve the patch foraging problem via an integration process within the frontal cortex, analogous to evidence accumulation in sensory decision making.

Results

A head-fixed patch foraging task for mice

We developed a head-fixed virtual patch foraging task for mice (Fig. 1), in which they ran on a cylindrical treadmill to traverse a one-dimensional virtual linear corridor. After mice ran a fixed distance, a visual "proximity cue" was presented, indicating the availability of a resource patch. If mice stopped running while the proximity cue was on, a patch trial was initiated with a visual cue, and water droplets were delivered stochastically while mice remained on the patch. Each patch delivered water droplets of one of three sizes (1, 2, or 4 μL), which was consistent within a patch. Rewards were delivered probabilistically each second on the patch, with probability given by a decaying exponential with a time constant $\tau = 8\ sec$ and one of three starting probabilities (0.125, 0.25, or 0.5). There were thus nine patch types: three reward sizes × three reward frequencies (Fig. 1B). To encourage staying, the first reward was delivered on 100% of trials (at t=0, the time of stop). At any time, mice could leave the patch by travelling a fixed distance on the wheel, at which point they would enter the inter-patch-interval and be required to run to the next patch to receive more water. Therefore, the task required mice to decide when to leave, given the reward history and time spent on the patch.

Patch wait times depend consistently on both reward size and frequency

During the foraging task, mice consistently guided their behavior towards obtaining water rewards. Example trials from two mice demonstrate how their running and licking evolved in response to task events (Fig. 1D), stopping in response to proximity cues and licking in expectation of rewards. Of the two mice, mouse 80 was less patient, often accelerating whenever rewards were not being delivered, whereas mouse 78 waited substantially longer in the absence of rewards. Reward seeking behavior was

evident across the population. In response to the proximity cue, mice successfully stopped to enter most patches (Fig. 1E). As mice stopped to enter a patch, they showed anticipatory licking at the same time as they began decreasing their speed (Fig. 1F), demonstrating an expectation of reward availability. After receiving an initial reward at t=0 (consistent across all patch trials), they began modulating their behavior in response to reward size. Lick rates were higher on patches with larger reward size, and mice waited longer before increasing their running speed (Fig. 1F).

The canonical signature of value-dependency during patch foraging is greater patch residence time (PRT) as a function of richer reward statistics. Put simply, efficient foragers should spend more time on patches with more abundant resources, as this increases their overall reward rate relative to indiscriminate foraging. Mice indeed showed higher PRT in patches with larger and/or more frequent rewards (example session, Fig. 1G; example mouse, Fig. 1H; all mice, Fig. 1I), with substantial variability in mean PRT across subjects. We therefore confirmed that mice performed the patch foraging task in a manner consistent with one of the fundamental tenets of foraging theory, albeit with marked differences in overall willingness to wait across subjects.

Reward sensitivity scales with level of patience across and within mice

To quantify the dependence of wait times on the size and frequency of rewards, we regressed PRTs on these factors and found that reward sensitivity varied substantially across the population. Notably, mice with higher overall PRT consistently yielded larger regression coefficients on reward size and frequency (Fig. 2A, Left, Right, $P<0.0001$ linear regression), demonstrating that more patient mice also displayed greater sensitivity to rewards. Coefficients for reward size and frequency covaried significantly (Fig. 2B, Left, $P<0.0001$ linear regression). Additionally, we observed an interaction between reward size and frequency (Fig. 2B, Right, $P=0.013$, ANOVA), suggesting that mice integrated these

117 factors together to guide their decisions, rather than using simple heuristics based on size and frequency
118 separately. We next compared regression models using different additive and multiplicative
119 combinations of these factors to determine which best accounted for behavior across patch types (Fig.
120 2C). We found that reward size explained PRTs across conditions better than reward frequency, with
121 regressions including the interaction of both factors performing best. Based on the observation that
122 mice's relative PRTs across patch types appeared approximately co-linear after log-scaling, we tested
123 whether a generalized linear mixed-effects model (GLME) could capture variability of PRT across mice
124 and patch types (Fig. S2). We fit a 3-factor GLME with fixed effects for reward size and frequency and
125 mouse ID as a random effect. We found that this captured PRTs across conditions for most mice (Fig.
126 S2C), further suggesting a common feature of reward sensitivity scaling across mice, despite the overall
127 differences in magnitude.

128 In addition to variability across mice, we observed substantial within-subject variability in PRT
129 over time. Some mice displayed striking fluctuations in their PRTs across sessions (Fig. 2D, Left: example
130 mouse. Fig. S1D, population), with the standard deviation in across session PRT ranging from 0.91 to
131 11.6s (Fig. 2D, Right). PRTs also varied within session, often gradually such that waiting times on
132 neighboring patches were correlated with one another (Fig. 2E). Taking advantage of these slow
133 fluctuations, we used PRTs from nearby patches as a proxy for estimating subjects' latent degree of
134 'patience' on a given trial (Fig. 2F, Left). We averaged surrounding PRTs with a Gaussian kernel (excluding
135 the current patch) to obtain a rolling estimate of patience across patches. We found that estimates of
136 patience took different shapes over sessions across and within subjects (Fig. 2F, Right), often with non-
137 monotonic and multi-phasic changes over time, thereby limiting the potential of capturing these effects
138 with simple regressions. Latent estimations of patience were highly predictive of PRT across trials (Fig.
139 2G, Left: example mouse, Right: population).

We then asked whether variability in patience within mice scaled with reward sensitivity similarly to the effects observed across mice. We assigned each trial to one of two groupings ("impulsive" or "patient") split by the median patience estimate per mouse. We observed that PRT scaling across the nine patch types was indeed steeper during patient trials (Fig. 2H, example mouse). To quantify this, we repeated our regression analyses from Fig. 2A separately for PRTs on the impulsive and patient groupings and found that regression coefficients for both size and frequency of rewards were consistently larger during the patient trials (Fig. 2I, $P<0.0001$ for size and frequency, Wilcoxon signed-rank test). These results demonstrate that within-session variations in patience do not simply add constant offsets to mouse PRTs but also scale reward sensitivity, similarly to how reward sensitivity scales with mean PRT across mice (Fig. 2A).

An integrate-to-bound process explains deviations from MVT predictions

The MVT dictates that, if patches are monotonically depleting in value as resources are consumed, an optimal forager should leave a patch once the instantaneous rate of reward drops below the average expected reward rate of the environment. Thus, irrespective of initial resource abundance, patches should all be left around this same threshold crossing (Fig. 3A, Left). To test this prediction of MVT, we computed expected reward rate at the time of each patch leave. We found that mouse behavior violated this prediction, with animals leaving high value patches too early, and low value patches too late, relative to an ideal observer (Fig. 3A, middle: example mouse, $P<0.0001$ 2-way ANOVA; right: population means, $P<0.0001$ linear mixed-effects model). This raises the question of whether mice may have incomplete task knowledge (see Discussion).

We next consider specific predictions of MVT reward rate estimation, focusing on the implications of exponentially decaying reward rates. Because all patch types had exponentially decaying

reward probabilities with the same time constant (Fig. 1C), differences in PRTs across patch types should remain constant, regardless of the threshold (Fig. 3B; see Methods). Furthermore, because we used patches with proportional reward sizes (1, 2, and 4µL), MVT predicts that the difference in leave time between 4 and 2 μL patches ($PRT_{4\mu L} - PRT_{2\mu L}$) should be equal to the difference between 2 and 1 μL patches ($PRT_{2\mu L} - PRT_{1\mu L}$; Fig. 3B; see Methods). Neither of these predictions was borne out in the data. Instead, reward sensitivity grew with mean PRT (Fig. 2A) and we consistently observed $PRT_{4\mu L} - PRT_{2\mu L} > PRT_{2\mu L} - PRT_{1\mu L}$ (Fig. 3C, P=0.0002, Wilcoxon signed-rank test).

In contrast, integrator models can explain the above empirical departures from MVT predictions. Parameters can be chosen such that the reward rate at leave time differs between patch types. Furthermore, in integrator models, differences in PRT between patch types necessarily grow with time, and slopes can be chosen such that $PRT_{4\mu L} - PRT_{2\mu L} > PRT_{2\mu L} - PRT_{1\mu L}$ (Fig. 3B, right). Notably, prior work has emphasized that integrator models can achieve MVT optimal behavior [9]; here, we show that while mouse wait times increase with patch value (a core feature of optimal foraging), they *deviate* from optimality in ways that can be explained by an integration process that is not perfectly tuned to reward statistics. Therefore, while we cannot rule out additional assumptions that may render our data consistent with MVT (e.g. imperfect task knowledge), we took integrator models to be the most parsimonious explanation of our behavioral data, and proceeded to fit integrator models to our data to narrow down the nature of the integration algorithm.

A competitive integration process explains patch leaving decisions when accounting for latent patience scaling

To approximate the generative process mice use to solve the patch leaving problem, we devised several integrator models which differentially track time and reward events, yielding probabilistic

predictions of patch leaving decisions. For each model, we compute a decision variable, *DV*, for each time bin, which is then transformed through a softmax function to generate a predicted probability, *P(Leave)*, of the subject leaving the patch (Fig. 3D – see Methods for detailed model description). As *DV* increases, so does *P(Leave)*. Because patch foraging is not a forced-choice process, but instead one that can be made continuously over time, we must remain agnostic over the temporal precision of the decision process. We therefore included a free parameter, P_{max}, constraining the maximal output value from the sigmoid. As a result, the midpoint of the sigmoid, corresponding to instances when the $DV = 0$, is not an indifference point wherein $P(Leave) = 0.5$, but instead is the point at which $P(Leave) = \frac{P_{max}}{2}$.

We fit three distinct integrator models for computing DVs over time (Fig. 3E). Common to each model is a slope at which the *DV* ramps up linearly over time in the absence of rewards, with its rate of increase scaled by reward size. We computed this reward size scaling as a power law function across reward sizes. What distinguishes the models is how they respond to reward deliveries. Model 1 exclusively tracks time on patch, and is indifferent to rewards, ramping upwards irrespective of additional reward deliveries. Model 2 is a full reset model, which returns to its initial baseline value upon each reward delivery. Model 2 is therefore "memoryless," in the sense that its value is dictated only by the time since the most recent reward and the patch's reward size. Model 3 performs a more sophisticated integration, incorporating rewards over time instead of only the most recent one. For each reward, Model 3 is reduced by an amount governed by a free parameter, *R*. For the purposes of this model, we fit only a single value for *R* rather than scaling it per µL, as we found that scaling slope alone was sufficient to account for the observed differences in PRT based on reward size. As we previously observed, mice's overall PRTs, as well as their reward sensitivity, were both strongly influenced by their latent levels of patience. We therefore took our previously obtained latent estimates of patience and allow them to scale both slope and P_{max} (Fig. 3F). We found that inclusion of this latent state scaling was

critical to accurately predict mice's patch leaving behavior across trials. Non-latent-scaled versions of each model performed substantially worse than their latent-scaled counterparts (Fig. S3B, Bonferroni-adjusted, P=0.0002, Wilcoxon sign-ranked test).

To better understand how rewards affect the decision process, we examined predicted leave probabilities from the respective latent-scaled integrator models. Model 1 can be ruled out a priori, as it could not account for the observed dependency of PRT on reward frequency, but we included it still as a baseline model for comparison. In contrast, Models 2 and 3 could both potentially account for the reward effects, as each produces greater PRT as a function of reward size and frequency by reducing their DV and P(Leave) predictions following reward deliveries. We therefore sought a systematic means of comparing model performance beyond predicting mean PRT, and instead focused on the probabilistic leave decisions the models predicted within individual trials.

To fit and compare the models, we performed maximum likelihood estimation to determine the best set of parameter fits for each subject. Iterating over one-second bins, we summed log probability of the models' output $P(Leave)$, adding $\log\left(1 - P_{i,t}(Leave)\right)$ for each bin mice remained on the patch, and $\log\left(P_{i,t}(Leave)\right)$ for bins when mice left the patch. The parameters that yielded the highest log probability for each mouse were then considered the best fit, and BIC values were calculated to assess fits for each model (Fig. 3G), while penalizing models for additional free parameters. BIC was lowest for Model 3, suggesting that a reward integrator best explains mouse leaving behavior across trials ($P<0.0001$ Model 3 over Model 1, $P = 0.0009$ Model 3 over 2, Wilcoxon signed-rank test, Bonferroni correction). Bayesian model selection further supported this result, confirming Model 3 had the highest protected exceedance probability, which measures the probability that a given model is represented more frequently than counterparts across the population.

We then took the Model 3 fits for each mouse and simulated PRTs, using *P(Leave)* outputs to probabilistically generate patch leaving decisions. Mean simulated PRTs per patch type from Model 3 closely matched mice's empirical PRTs across the population (Fig. 3H, MSE = 0.413s). We further corroborated the model fits cross-validation and obtained similar results. For each fold, we used maximum likelihood estimation to obtain parameter fits across the four training folds. Those parameter fits were then used to generate predicted *P(Leave)* across bins in the held-out test fold patches, over which we summed log predictive probability of patch leaving decisions. Results from the cross-validation were nearly identical to the original fits, with log probability of model predictions consistently higher for Model 3 than the alternative models (Fig. S3E, P<0.0001 Model 3 over Model 1, P = 0.0008 Model 3 over 2, Wilcoxon signed-rank test, Bonferroni correction). When comparing the mean PRT across patch types from simulations with mice's corresponding PRT on held-out test trials, we observed that both Model 2 and Model 3 matched mice's aggregate behavior well (Fig. S3F).

We next asked how well we could predict PRT across individual trials using the Model 3 parameter fits from the training folds (Fig. 3I, example trial). For each patch, we calculated the leave density per one-second time bin by multiplying the fraction of trials the model predicted to still be on the patch at the start of the time bin by that time bin's predicted leave probability: $Fraction\ on\ patch_{t,i} * P_{t,i}(Leave) = Leave\ density_{t,i}$. We found that model predicted PRTs explained a substantial amount of the variance across individual patches (Fig 3J, Left: population, median R^2 = .54, Right: example mouse, R^2 = .80). An example session is shown to demonstrate how these model predictions capture PRT variance beyond that which is predicted by the latent patience estimates (Fig. 3K, Fig. S3I).

To further distinguish between Models 2 and 3, we took advantage of the richness in trial history that the task provides and examined trial types with specific histories for which Models 2 and 3 make divergent predictions (Fig. 3L). We compared two combinations of reward sequences: 'RRR' trials, which

delivered rewards at t=0,1,2 seconds, and 'R0R' trials, which delivered rewards t=0 and 2 seconds (but no reward at t=1). Because Model 2 produces a full reset following reward, it predicts identical leaving behavior on both trial types, as both reset to the same value after the final reward (Fig. 3L, left). In contrast, Model 3 integrates reward history and preserves the effect of the t=1 reward, predicting higher PRT on 'RRR' over 'R0R' patches (Fig. 3L, right). Simulations demonstrate these differences (Fig. 3M, left, middle), and mouse behavior strongly matched Model 3 predictions, with nearly all subjects waiting longer on 'RRR' trials (Fig. 3M, right). These results support Model 3's reward integration over Model 2's reward reset as a better description of the decision algorithm. We similarly compared 'RR0' versus 'R0R' trials and found that mice behavior again better matched Model 3 predictions (Fig. S3G,H). In addition to our separate per mouse model fits, we fit behavioral models wherein a single set of free parameters was fit across mice. With only their respective latent estimates of patience across trials to scale the model predictions, we were still able to robustly capture the variability across mice and patch types (Fig. S3J, MSE = 0.988).

Slow ramping and rapid reward transients are prominent features of frontal cortex activity

We next aimed to investigate the neural underpinnings of the foraging strategies identified above. We recorded extracellular activity with Neuropixels probes broadly throughout frontal cortex and underlying subcortical areas while mice performed the task (Fig. 4A,B; N = 6090 units from 33 recording sessions in 9 mice). Peri-stimulus time histograms (PSTHs) aligned to patch stop revealed units whose activity ramped upward as mice remained in the patch, with ramp slopes increasing with reward size (example in Fig. 4C). For this example unit, firing rate ramped to a common threshold prior to patch leave, was inhibited by reward delivery, and showed ramping activity spanning up to tens of seconds within individual patches (Fig. 4C, bottom)—much like the ramping decision variables identified through

behavioral modeling. In addition to these "ramping neurons," we observed a diverse array of activity patterns, as expected in a complex task, including "reward-responsive" neurons which responded transiently to reward delivery (Fig. 4D).

Consistent with the presence of ramping activity in our neural data, the activity of individual neurons was significantly correlated with ramping decision variables derived from behavioral models (Fig. 4E). Correlations could be either positive or negative but were more frequently positive than negative (1308 and 881 cells out of 6090 with z-test $P < 0.001$ vs shuffle, respectively; Fig. 4E). Consistent with this, the top principal component(s) of neural activity on patches often resembled an upward ramp which was decremented by reward (Fig. 4G). Notably, greater variance in single neurons' firing rates was explained by the model 3 decision variable in frontal cortex compared to subcortical areas, suggesting a specialized role for frontal cortex in this task (Fig. 4F). However, decision variables were also correlated with behaviors such as running speed and lick rate (Fig. S4A). These correlations were generally lower in longer compared to shorter patches but remained non-zero (Fig. S4B). Such correlations may reflect embodied cognition and therefore should not necessarily be viewed as confounds per se (see Discussion). Nevertheless, to rigorously isolate task-related activity from correlated behavioral variables, we next turned to a multivariable regression approach using generalized linear models.

GLM modeling with unsupervised clustering reveals six clusters of neurons, the most prominent of which shows ramping activity and reward responses with opposite signs

To comprehensively characterize the space of task-related neural activity in our task, we used a Poisson Generalized Linear Model (Poisson GLM; Fig. 5A; snippet of GLM fit to an example neuron, Fig. 5B) which included the following regressors: session time and its square, to capture slowly varying changes in firing rate; behavioral variables derived from the rotation of the running wheel (position,

speed, and acceleration) and the lick sensor (lick rate and its derivative); discrete events at patch stop, patch leave, and reward times, which were convolved with a raised cosine basis, separately for each reward size; and three ramping "decision variables" which could be used to construct the decision variables in the behavioral models (Time on Patch, Total Reward, and Time Since Reward), separately per reward size. In total, the GLM included 85 variables. Elastic net regularization with 90% L1 (regularization strength selected through cross-validation, see Methods) encouraged the model towards sparsity, thus mitigating issues related to correlated regressors. As a further test of whether behavioral variables such as running speed may contribute to ramping activity in our data, we fit also GLMs to inter-trial intervals (ITIs), and found that GLM coefficients for running speed were more correlated between odd and even patches than between odd and even ITIs, suggesting that even if ramping activity is correlated with speed it is not a low-level, non-specific motor code that is common to both patches and ITIs (Fig. S5).

To assess model performance, we computed cross-validated percent deviance explained compared to a null model (Fig. 5C, top; mean % deviance explained $\pm$ SEM = 5.2 $\pm$ 0.1, N = 6090 neurons). Reward kernels and decision variables were termed "Task Variables" and used to identify task-related neurons. Percent deviance explained was computed for full models compared to models without Task Variables ("reduced models"), and neurons with deviance explained > 1% in this comparison were deemed "task-related" (Fig. 5C, bottom; 1458/6090 neurons, mean % deviance explained versus reduced model $\pm$ SEM = 0.87 $\pm$ 0.03). GLM performance was higher for frontal cortex areas than other areas (Fig. 5D, mean % deviance explained $\pm$ SEM, Frontal Cortex = 6.5 $\pm$ 1.0, Subcortical Areas = 4.1 $\pm$ 0.6, paired t-test P = 0.016, N = 9 mice), indicating that frontal cortex may be especially relevant for our task. Aligning the activity of task-related neurons to patch stop, splitting trials by those in which a reward was delivered at 1 second or not (RR versus R0 trials), and sorting by the time of peak response revealed a "wave" of activity following each reward: A rapid reward response with variable latency which gradually transitioned into upward ramping activity (Fig. 5E), as exemplified by the two neurons shown in Figure 4.

We clustered task-related neurons based on their GLM coefficients to characterize the space of activity patterns in our task in an unbiased manner (Fig. 6). We additionally excluded 60 neurons whose coefficients on task variables were all zero, leaving 1398 task-related neurons in this analysis. The data to be clustered thus consisted of a matrix of 42 task-variable coefficients x 1398 task-related neurons (Fig. 6B). First, we reduced the dimensionality of this matrix with PCA, keeping the first 3 principal components (34.7% of variance; Fig. 6C). The top component (14.5% of variance) had reward kernel coefficients and ramping coefficients with opposite signs, indicating that ramping suppressed by rewards was a prevalent feature of task activity. We then clustered the dimensionality-reduced matrix using a Gaussian Mixture Model and K (the number of clusters) spanning 1-10. K=6 minimized the BIC and was chosen as the number of clusters (Fig. 6D).

The resulting 6 clusters could be grouped into 3 pairs with similar shapes but opposite signs (Fig. 6E,F, Fig. S6A): (1) A pair of clusters with a slow negative/positive reward response, scaled by reward size, and positive/negative coefficients on Time Since Reward (Clusters 1 and 2), (2) A pair of clusters with rapid reward response, scaled by reward size, and minimal ramping coefficients (Clusters 3 and 4), and (3) a pair of clusters with minimal reward coefficients but significant positive/negative coefficients on Time On Patch (Clusters 5 and 6). Similar clusters could be identified in 6/9 mice; the other 3 mice had too few task-relevant neurons to reliably identify the same patterns (Fig. S6B). PSTHs of the average activity of each cluster aligned to patch stop and split by reward delivery at one second matched the patterns of GLM coefficients (Fig. 6G). Moreover, splitting trials by PRT or aligning to patch leave revealed that Cluster 1, and to a lesser extent Cluster 2, ramped to a common threshold prior to patch leave, whereas Clusters 5 and 6 continued to ramp up/down as the animal stayed longer on the patch (Fig. 6H, Fig. S6C). Because of the clear ramp-to-threshold-like properties of Cluster 1, we considered this to be our “ramping” population (568/1398 task-relevant neurons).

We examined how these clusters mapped on to brain regions recorded in our task. A greater percentage of neurons in frontal cortex areas was identified as "task-related" compared to other areas (Fig. 6I; frontal cortex areas: 916/3156 [29%], other areas: 482/2934 [16%]), consistent with a specialized role for frontal cortex in our task. However, of task-related neurons, similar fractions were assigned to Cluster 1 across brain areas (Frontal Cortex: 40.7%, Subcortical Areas: 40.5%).

While our recordings consisted of untagged populations of neurons, thus mixing diverse cell types, spike waveform can be a marker for certain cell types in the brain [15, 16]. To assess whether our GMM Clusters may be enriched for certain cell types, we computed spike waveform for each neuron in our dataset and classified the waveforms as either Regular or Narrow based on spike width (Fig. S6D-F). In Frontal Cortex recordings, we observed that GMM Cluster 3 (transient positive response to reward) contained an over-representation of Narrow waveforms (Fig. 6J). This indicates that our functionally defined clusters may map onto cell type-specific populations of neurons in the cortex, a possible direction for future study.

Decision variables derived from behavioral modeling are represented in frontal cortex dynamics

To test whether the latent decision variables identified through behavioral model fitting were represented in neural activity, we fit cross-validated linear models predicting each latent variable from the joint activity of co-recorded neurons (Fig. 7A). Decision variables could be reliably decoded from neural activity in most sessions (Fig. 7B,C). Although Model 3 best explained mouse behavior (Fig. 3), decision variables from Model 2 could be decoded from neural activity with similar fidelity, and Model 1 with only slightly lower fidelity, consistent with a recent report [14] (mean CV $R^2 \pm$ SEM for Model 1 = 0.31 ± 0.03, Model 2 = 0.38 ± 0.04, Model 3 = 0.35 ± 0.03; N = 25 sessions after excluding 5 sessions

with CV R^2 < -0.2, see Methods; Fig. 7D). Moreover, Model 3 decision variable decoding from neural activity was unrelated to performance of the behavioral fit on a per-session basis (Fig. S7).

To assess how the ability to decode the Model 3 decision variable relates to brain region, we chose sessions in which we had sufficient neurons in both Frontal Cortex and Subcortical Areas (least 20 neurons each). Units were down-sampled so that both areas contained the same number of neurons in this analysis. In this within-session, unit-number-matched comparison, decoding for Model 3 DV was better in Frontal Cortex compared to Subcortical Areas (Fig. 7E; mean CV $R^2 \pm$ SEM, Frontal Cortex = 0.34 $\pm$ 0.03, Subcortical Areas = 0.28 $\pm$ 0.03, paired t-test P = 0.0037, N = 23 sessions after excluding 7 sessions with insufficient neurons and 3 sessions with CV R^2 < -0.2 for both frontal cortex and subcortical fits).

State space modeling demonstrates that neural population activity is ramping, not stepping, on single trials

In previous sections, we identified a cluster of neurons that showed ramping activity on average ("Cluster 1", Fig. 6). We next tested whether single trial dynamics within this sub-population (e.g. Fig. 8A) were also continuously varying ramps or were discrete processes [17, 18]. We fit a continuous ramping model and a discrete state stepping model to the neural population response (Cluster 1 spike rate, summed across neurons; Fig. 8B,C). The ramping model had a pulse response to rewards and otherwise ramped in the absence of rewards. The stepping model had two discrete firing rate states; the firing rate started in the low state and probabilistically jumped to the higher state following Markovian dynamics. Additionally, to capture reward responses, the firing rate returned to the low state after rewards, whereas otherwise the firing rate was fixed to remain in the high state. We found that 16/16 sessions with at least 10 Cluster 1 neurons were better fit by the ramping than the stepping model (Fig. 8D).

Additionally, the impact of rewards on the latent accumulator was more negative for larger reward sizes (Fig. 8E, left; regression slope of model parameter vs. reward size, mean $\pm$ SEM = -0.012 $\pm$ 0.0041, t-test P = 0.0093, N = 16 sessions), and slopes were marginally shallower (Fig. 8D, right; regression slope of model parameter vs. reward size, mean $\pm$ SEM = -0.025 $\pm$ 0.012, t-test P = 0.065, N = 16 sessions), consistent with the hypothesized integration mechanism which accounts for the impact of rewards on mouse wait times (Fig. 3).

Discussion

Here, we present a novel virtual reality-based patch foraging task for mice. We varied patch richness along two independent dimensions (reward size and frequency) and demonstrated that mouse patch residence times were sensitive to both, matching a basic prediction of the MVT. However, closer examination of mouse behavior revealed discrepancies with the theory: The instantaneous reward rate at leave time differed between patch types, and differences in PRT between patch types increased with mean PRT. While we cannot rule out alternative formulations of MVT that could capture these effects, we found that both are explained by integrator models. We fit several variants of integrator models (Reward Indifferent, Reward Reset, and Reward Integrator) and identified the Reward Integrator (Model 3) as the best fit to our behavioral data, using both quantitative model comparison and diagnostic trial type comparisons. Decision variables from the Reward Integrator model were represented in neural activity more robustly in Frontal Cortex compared to subcortical areas. Using regression analysis and clustering, we identified a subset of "ramping" neurons (Cluster 1) and found that the activity of these neurons was better described by continuous ramping rather than a discrete stepping process. Together, these results suggest that a continuous integration process reflected in the activity of frontal cortex neurons is a likely mechanism by which mice make patch leaving decisions during foraging.

Reward sensitivity scaling across varying levels of patience

Mice displayed consistent value-sensitive behavior in our experiments, spending more time in patches in which rewards were larger and occurred more frequently. Looking at the effects of these reward factors more closely, we observed that PRT dependence on reward size was greater than reward frequency. This result is logical because reward size was an explicit distinction across patch types – if mice were on a patch with large rewards, every reward delivered was 4µL. In contrast, reward frequency

425 was a hidden feature per patch type, governing the probabilities used to generate the stochastic
426 rewards, but not explicitly controlling their outcome. Consequently, some high frequency patches, by
427 chance, could yield fewer reward deliveries than some lower frequency patches.

428 Decision making varies dramatically with the internal state of the organism [19] [3] [20] [21].
429 While most mice were value-based in their PRT behavior, they exhibited substantial variability in their
430 overall willingness to wait, which we describe as "patience." Interestingly, we observed that reward
431 sensitivity scaled systematically with variability in patience, resulting in greater differences in PRT as a
432 function of patch value when patience was higher. Coefficients for reward size and frequency effects on
433 PRT varied widely, but they did so with approximately the same ratio across mice. Because of this
434 common scaling, we were able to consistently capture PRTs across conditions and individuals with a
435 GLME where reward effects were fixed across the population, and with only a random effect term for
436 mouse ID to scale as a multiplier. These results suggest that mice with different PRT behavior may use a
437 common algorithm, with individual differences in scaling. Consistent with this hypothesis, we were able
438 to robustly predict PRTs across the population with our generative integrator model using a single set of
439 free parameters across the population, with the only source of individual differences being achieved via
440 the scaling introduced by the per patch estimates of patience.

441

442 *Contrasting results with the Marginal Value Theorem*

443 Mouse behavior qualitatively matched one of the fundamental predictions of the MVT: they
444 spent more time foraging in patches where water resources were more plentiful. However, when
445 examining quantitative predictions made by MVT, we found inconsistencies with our empirical data.
446 When comparing instantaneous expected reward rate at time of patch leaving between mice and an
447 ideal observer MVT model, we found that all mice significantly deviated from MVT predictions for

reward size, and the majority also did for reward frequency. However, expecting mice to adhere perfectly to idealized MVT behavior makes unrealistic demands on their knowledge of task structure. We have no reason to expect that mice fully understand the task structure – for example, that there are nine discrete patch types and that reward size always remains the same within a given patch. It would require many observations before mice could plausibly learn these nuances, and the fact that many mice perform the task in a reward-sensitive manner even on the first day of training indicates use of a simpler algorithm. Moreover, it would not be an effective strategy in naturalistic environments to overfit behavior to the statistics of patches so precisely, as environmental statistics are constantly changing and animals need to be able to flexibly adjust their behavior to those changes, as well as be able to forage effectively in novel environments.

The central test of MVT is to compare instantaneous reward rate at the time of patch leave with the average reward rate in the environment. We do not employ this analysis for several reasons. Because we did not have an experimental condition in which we manipulate environmental value, we would not have sufficient control over this variable for rigorous analysis. Changes in environmental value in our task only occur due to either 1) temporary fluctuations due to task stochasticity; or 2) changes in the animals' own behavior. The former are too short-lived and infrequent to provide sufficient analytical power, and the latter would be inexorably confounded by circular logic if we attempted to analyze changes in behavior due to value differences while selecting for periods when the value differences were caused by a change in behavior. Moreover, making this kind of comparison rests on the assumption that we could comprehensively quantify environment value. In reality, we are only privy to variables that we can directly measure or control, as well as those that can be rigorously inferred through other correlations. Defining reward simply as water collected is convenient for certain analyses, but it is not comprehensive. Animals are not strictly reward maximizers. They are guided more generally by utility, which can be influenced by many factors (e.g. fear of predation, energy levels, competition), the majority of which are

beyond our ability to measure. One such confound we observed in our experiments was the value of running. Intuitively, this would seem to be a cost since it requires time and energy expenditure to get from one patch to the next. However, we observed that mice, despite being water-restricted, would sometimes skip substantial numbers of patches in a row, choosing instead to run at high speed. This behavior indicates a positive utility placed on running, sufficiently so to forego immediately available rewards, even when mice were value-based on the patches they did engage. Furthermore, the variability we observed in patience within and across mice further emphasizes that we cannot make any precise claims about overall value.

While previous studies have demonstrated effects such as animals overstaying in patches beyond that which would be predicted by MVT, these can often be explained by other factors influencing the relative value comparison between patch and environmental value [22-24]. Despite these limitations, we identified a means of testing MVT predictions that is robust to experimenter uncertainty about subjective overall value. By instead focusing our analyses on the relative PRTs across patches with different reward sizes, we could overcome this epistemic limitation because irrespective of the overall value mice place on the environment, relative reward rates across patches of different reward size are unaffected. Using this framework, we observed that mice did not appear to be using a pure reward rate estimation strategy, as MVT predicts. Instead, we found that mouse behavior could be explained by a competitive integration process, and that deviations we observed from MVT emerge as a direct consequence of an integration algorithm. Moreover, we found that dynamics matching predictions of an integration process were a prominent feature throughout our recordings in frontal cortex. While we cannot rule out the possibility that further modifications could be added to MVT to account for the deviations we observed, our results build on previous theoretical work and demonstrate the viability of integrate-to-bound dynamics as a parsimonious explanation for how mice, and potentially other animals, solve the patch leaving problem.

A competitive integration model as an algorithmic solution to the patch leaving problem

We constructed several different integrator models aimed at capturing the algorithm mice use to solve the patch leaving problem. Instead of devising them as pure integrate-to-bound models with a deterministic threshold, we formulated them as cognitive models of an ongoing decision process. To account for observed differences in PRT, we allowed integration slope to scale by patch reward size. In addition, allowing the models to scale by relative levels of patience was critical to achieving quality fits on mice behavior, so much so in fact that an alternative model which ignored reward effects entirely provided a better fit than Model 3 stripped of its patience scaling. In addition to patience scaling allowing the model to capture variability in PRT across trials within mice, we fit models using a single set of parameters across all mice, allowing only patience scaling to produce differences. We found that patience scaling over this population captured PRTs across mice, demonstrating that latent patience scaling of a common algorithm could account for PRT differences both within and across mice.

Model selection criteria identified Model 3 as the superior fit, but we also found that the reward reset Model 2 could account for a large fraction of the variance in mice PRT. Other studies have demonstrated mice who were significantly better matched to a reward reset process over a reward integration strategy, but that behavior only shifted from a reward integration strategy after sufficient training on a task in which reward reset was the optimal strategy [13]. Other experiments have found a combination of the two strategies best explained mice behavior [25]. We do not rule out the possibility of reward reset effects influencing mice behavior in our experiments, but we do show positive signs of Model 3 that could not be explained by a full reset model. We also do not rule out a model which combines features of both. Further work is required to more closely inspect relative influences of the two strategies.

519 Our behavioral modeling builds on previous theoretical work and provides a framework for
520 future investigation. We present evidence in support of an algorithm in which antagonistic effects of
521 time and reward are integrated to modulate patch leaving decisions such that greater time is spent on
522 patches in which greater amounts of reward is delivered. The model accurately predicted leaving
523 probabilities across time bins and was able to predict PRT across single trials with remarkable precision.

524

525 *Regional specificity of the computation*

526 Previous studies have demonstrated that perturbation of a number of rodent frontal cortex
527 regions affects waiting times [12, 26, 27] [28]. Our recordings point to frontal cortex broadly
528 (encompassing OFC, ACC, PL, IL, M2 and M1) as an area with enriched task-related activity. Specifically,
529 ramping decision variables derived from behavioral models could be better decoded from neural activity
530 in frontal cortex compared to subcortical areas (Fig. 7E). Our observational study therefore sets the
531 stage for future experiments using optogenetic perturbations to test the functional contributions of
532 various frontal areas and molecularly/connectivity-defined cell populations in a value-based foraging
533 task, as well as their precise timescales [29] [30].

534

535 *The possible neural basis of a "patience" latent state*

536 In our task, mouse PRTs fluctuated slowly throughout individual sessions with a time constant of
537 roughly ten trials, independently from the richness of individual patches. We interpret this as reflecting
538 an internal state which we describe as "willingness to wait" or "patience." This may be related to "active
539 persistence" which was causally linked to the activity of serotonergic neurons in a foraging task in a
540 recent study [31], although in that study activation of serotonergic neurons promoted actively continuing

to nose-poke for reward, whereas here patience reflects a passive process of waiting for rewards [32]. In our case, incorporating the "patience" state into our algorithmic behavioral models greatly improved model fits. This was accomplished by allowing patience to modulate model parameters: specifically, ramping slopes were divided by a power law scaling of each trial's relative latent value. Further work is needed to determine whether this is the way in which the brain incorporates the patience state into its decision-making process, and if so, what is the neural basis of this modulation.

Relationship between ramping neural activity and movement

As shown in Figure S4, several behavioral variables (most prominently position, speed, lick rate) correlated either positively or negatively with ramping decision variables, raising the possibility that some of the ramping activity in our neural data could be associated with task-related movements. Indeed, several recent studies have reported movement-related activity widely throughout the brain, beyond classically movement-related areas [33] [34]. Such correlations likely contribute to the ramping signals in our data, especially for mice in which these correlations were high. However, we believe movements cannot entirely explain ramping activity, for the following reasons: (1) Ramps were observed even in trials in which mice exhibited no movement by these metrics (Fig. 8A), (2) A population of ramping neurons ("Cluster 1") could be identified from GLM regression coefficients, in which movement-related activity had been regressed out by including behavioral variables in the model (Fig. 5, 6), (3) Cluster 1 firing rates preceding leave time were similar across reward sizes despite different leaving speeds (Fig. S6C, Fig. 1F), and (4) Speed coding was qualitatively distinct in the trial periods and inter-trial-intervals (Fig. S5). An additional possibility is that correlations between decision variables and behavior could, in part, reflect an embodied cognitive process [35]. The extent to which behavior is itself an element of natural decision-making represents an interesting direction for future study.

Limitations of using neural data to identify the behavioral strategy

Although Model 3 best explained mouse behavior in our task, decision variables from all three models could be decoded from neural activity with similar accuracy (Fig. 7D). This could reflect a "reservoir" of decision variables in frontal cortex, in which strategies that are not currently used are nevertheless continuously tracked in neural dynamics [14]. However, caution must be taken to consider the contribution of task-related movements to these results. For example, licking patterns are similar following each reward and may contribute more to a "Model 2" (reset)-like pattern of neural activity, regardless of the actual decision algorithm implemented by the brain. If we simply exclude *all* movement-related activity (which is not practical), then we risk throwing away activity related to embodied cognitive processes. At the same time, it is circular to identify movements as "embodied cognition" because they match a particular model, and then exclude other movement-related activity and draw conclusions based on the match between activity and behavior. Thus, we believe the neural data are of limited use in identifying the exact strategy employed by animals in this task. Instead, our argument for Model 3 is behavioral (Fig. 3), and we show that activity related to the decision variable from this model is *present* in neural activity (Fig. 4E, Fig. 7), providing a potential substrate for patch foraging calculations in the brain.

Ramping activity during decision making

We observed ramping spiking activity to be a prominent feature of frontal cortex dynamics as time elapsed while mice foraged in virtual patches. Ramps scaled with patch value, with shallower ramps on patches when rewards were larger. While the accumulation of time resulted in the upward progression of ramps, we observed that additional reward deliveries had an antagonistic effect,

587 inhibiting spiking activity and thereby driving the ramps further away from their terminal firing rates.
588 Together, these effects provide sufficient components for allocating more time to foraging in patches in
589 which rewards were more abundant. Building on theoretical work demonstrating reward-modulated
590 ramping activity to be a viable mechanistic solution to the patch leaving problem [9], we found that
591 these ramps were well-matched to decision variables in integrator models that robustly captured mice
592 PRT behavior and explained deviations from predictions of MVT. Together, these results point to
593 integrator dynamics as a parsimonious candidate solution to the patch leaving problem.

594 Ramping activity has been shown to be a prevalent feature of perceptual decision making across
595 a range of organisms and brain areas [36-40]. It provides a mechanism for making relative value
596 decisions [41], and precedes the onset of timed actions [42] [43] [44]. We observed similar ramping
597 dynamics within the context of naturalistic foraging. Moreover, we observed that single trial ramps were
598 apparent in population spiking activity on scales of up to tens of seconds, with reward-triggered
599 inhibition of those ramps followed by mice increasing their PRT before leaving patches. The flexibility of
600 ramps to scale their slope and integrate inputs with opposing valences, taken in the broader context,
601 suggests that ramping activity could be a ubiquitous feature of decision processes that unfold over time.

602

603

Main Figures

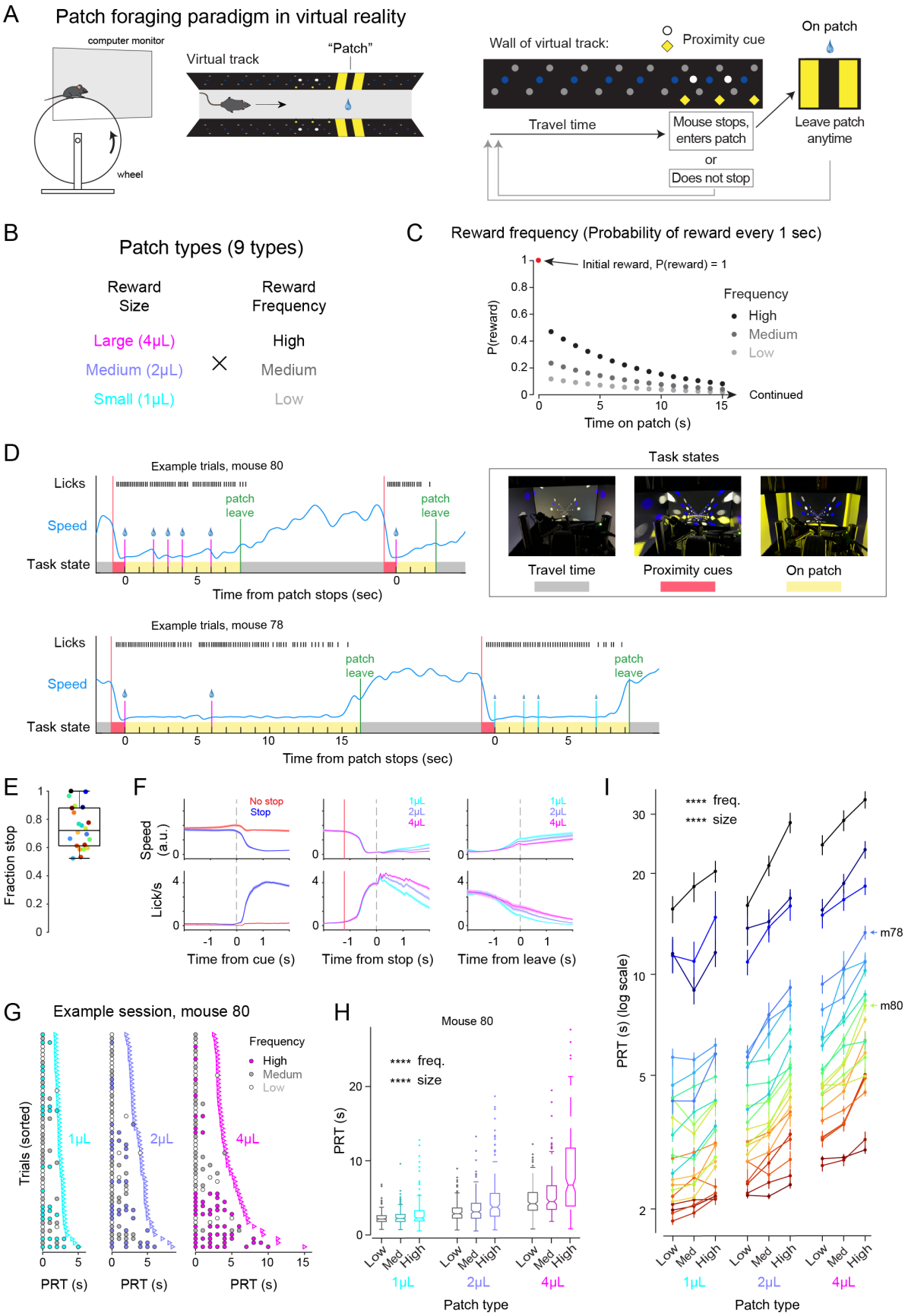


Figure 1: In a head-fixed patch foraging task, mice show consistent dependence of wait times on reward size and frequency.

A) Virtual linear track for patch foraging task.
B) Combinations of three reward sizes and frequencies yield nine patch types.
C) Probability of reward delivery after each 1 second interval per frequency condition.
D) Example trials from two mice. Mice stop running in response to the proximity cue to enter patches and receive stochastically delivered water rewards. Photographs of the three task states are shown. Monitor brightness was increased to maximum for the photos.
E) Mice stopping behavior in response to the proximity cue.
F) Mean population running speed (Top) and lick rate (Bottom) in response to task events. (Left) Aligned to proximity cues appearance, split by trials in which mice successfully stopped (blue) or skipped (red) the patch. (Middle) Aligned to patch stop, split by reward size. Red line indicates mean time of proximity cue appearance. (Right) Aligned to time mice left the patch, split by reward size.
G) Reward deliveries and patch leave times from an example session grouped by patch reward size. Trials are sorted in ascending order of PRT from top to bottom per reward size. Dots are shaded based on underlying reward frequency condition.
H) PRT per patch type from an example mouse. Column groupings separate patches by reward size. Within columns, patches are split by reward frequency.
I) Mean PRT per mouse across patch types, y-axis log-scaled. Colored by mouse. Same patch type groupings as H. Error bars indicate standard error of the mean.

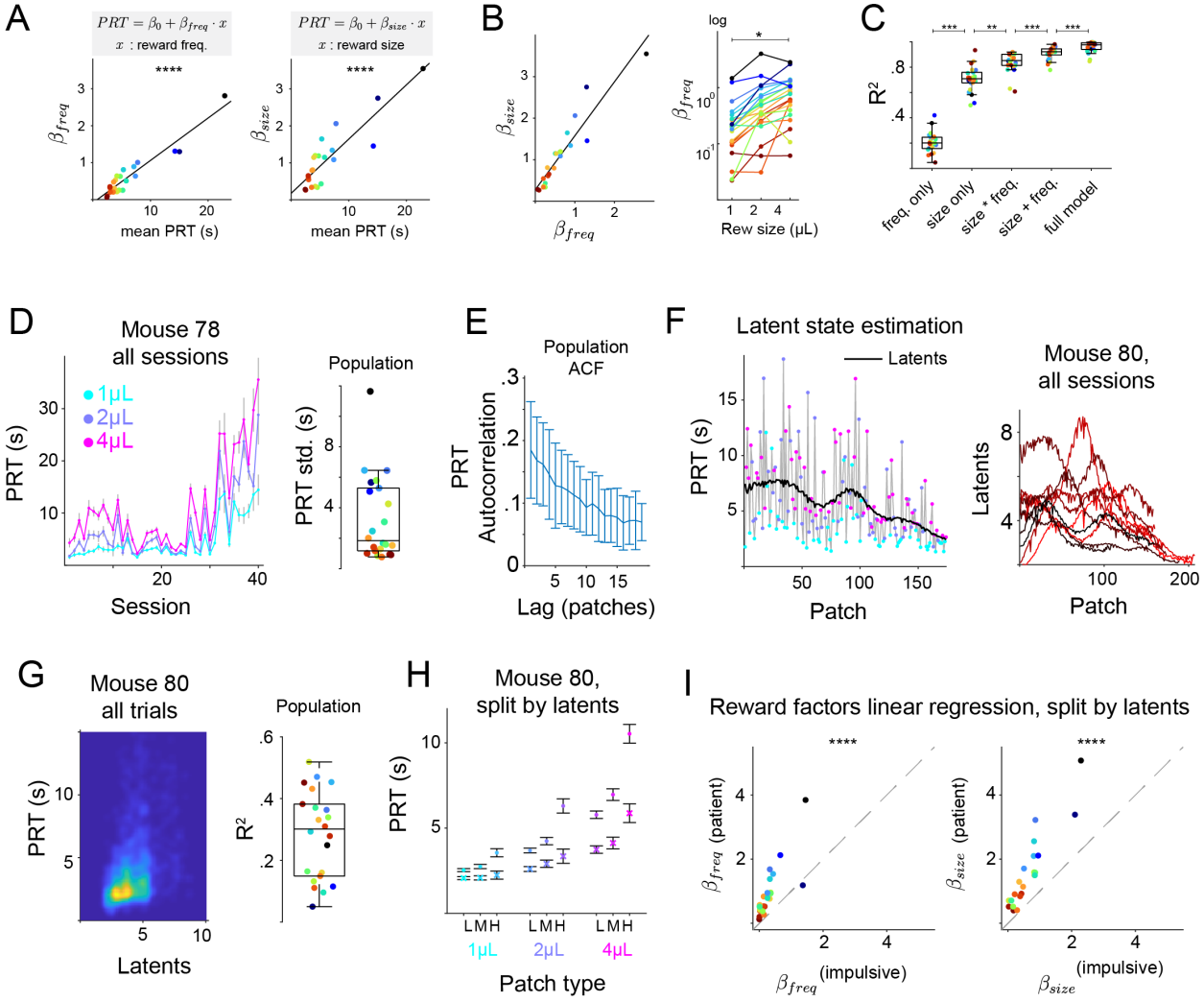


Figure 2: Reward sensitivity scales with patience across and within mice.

A) Linear regressions of PRT on reward factors per mouse. Left: Regression coefficient for PRT on reward frequency scales with mean PRT. Right: Same, for reward size.

B) Left: Regression coefficients for PRT on reward size and frequency scale together across mice. Right: Regression coefficients for PRT on reward frequency are larger for larger reward sizes.

C) R^2 values for predicting mean PRT per patch type across mice for different additive and/or multiplicative combinations of reward size and frequency.

D) Left: Mean PRT per µL across sessions, example mouse. Right: Standard deviation of mean PRT across sessions per subject.

E) Autocorrelation function of PRT over patches, mean across subjects. Error bars indicate standard deviation over per subject fits.

F) Gaussian smoothing over neighboring trials provides an estimate of latent state. Left: Example session latent state inference. Gray trace tracks PRT across patches with colored dots indicating reward size for each patch. Black trace indicates value from the latent state estimation. Right: Estimated latent across trials over all 10 sessions from mouse 01.

G) Latent state estimation accounts for variability in PRT across trials. Left: Smoothed heatmap of PRT versus estimated latents for all trials, example mouse. Right: Population R^2 values from regressions of PRT on estimated latents.

H) PRT per patch type with trials split into two latent state groupings: impulsive and patient, separated by median latent state, example mouse.

I) Regressions on reward factors, fit separately per latent state grouping across subjects.

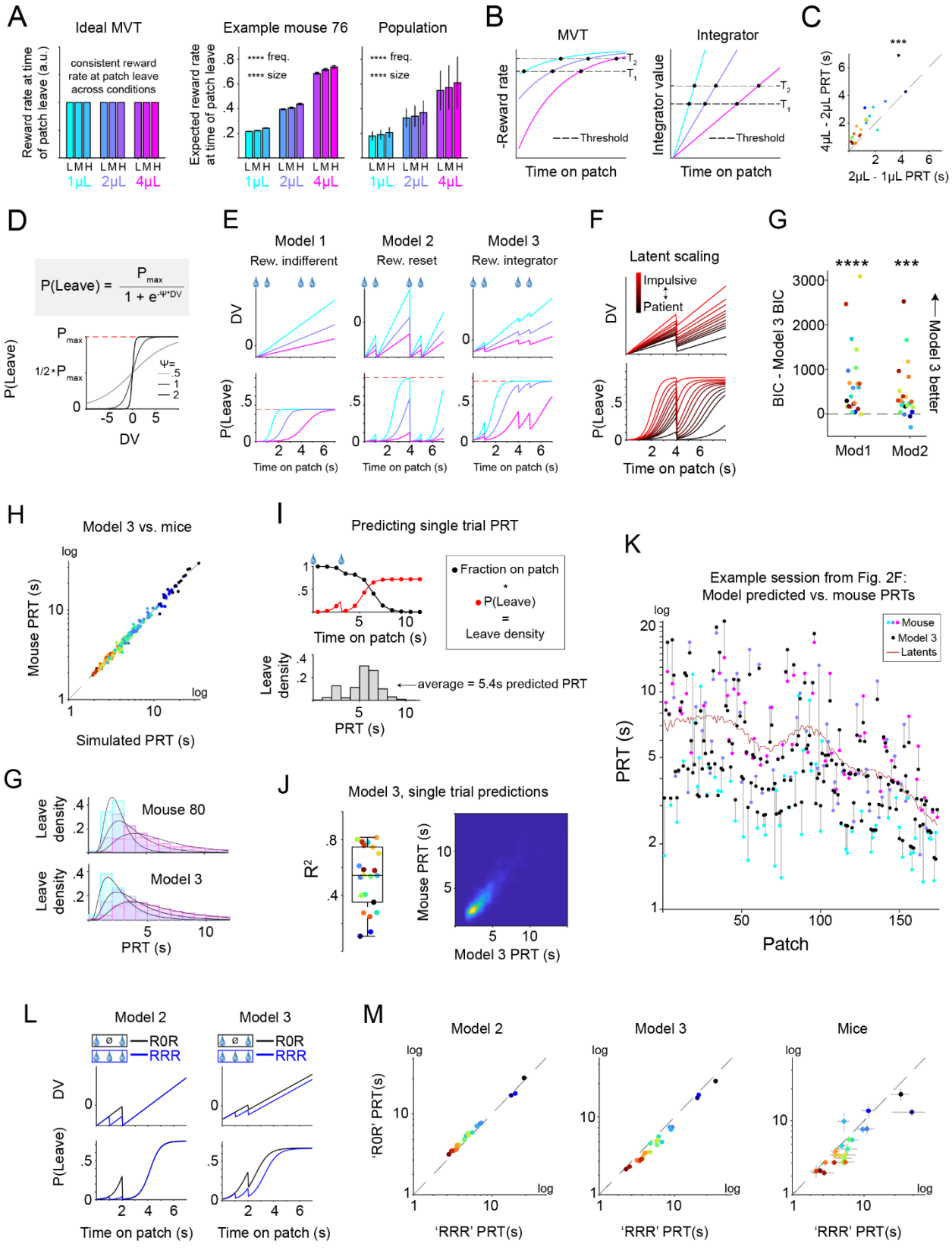


Figure 3: A competitive integration process accounts for violations of MVT predictions and explains mice behavior across patches

A) Instantaneous expected reward rate at time of patch leave across patch types, predicted by optimal MVT (Left), sample mouse (Middle, Bonferroni adjusted p<.0001 for size and frequency, 2-way ANOVA), and population (Right, p<.0001 for size and frequency, linear mixed-effects model). Error bars in for sample mouse indicate standard error of the mean. Error bars for population average indicate standard deviation across mice.

B) Schematic demonstrating how MVT predicts a constant difference in PRT across reward sizes irrespective of threshold (Left), whereas an integrate-to-threshold model predicts greater PRT differences across reward sizes for high thresholds (Right). Traces are colored by reward size. Dashed lines show two sample thresholds. Black dots indicate threshold crossings.

C) Mice show greater differences in mean PRT between 4µL-2µL patches, compared to 2µL-1µL patches.

D) Sigmoid transformation of decision variable (DV) into probability of mice leaving patches per one second interval for different inverse temperature values (light grey = 0.5, dark grey = 1.0, black = 2.0). Red dashed line indicates the maximum P(Leave) output, P_{max}.

E) Example schematics for DV (Top) and corresponding P(Leave) output (Bottom) for three different integrator models over patches with rewards delivered at t = [0,1,4,5] seconds. Decision variables (DV) ramp upwards over time. Model 1 does not respond to reward deliveries. Model 2 resets to its baseline value following any reward delivery. Model 3 integrates rewards with a constant negative value. Red dashed line indicates the maximum P(Leave) output, P_{max}.

F) Schematic demonstrating how DV and P(Leave) scale relative to latent patience estimation across patches, for a sample patch in which rewards were delivered 0 and 4 seconds. Black traces indicate patches with more patient estimations and ramp up more slowly. Red traces indicate impulsive latents and yield sharper ramps.

G) Relative BIC values for the model fits across subjects. Model 3 yields a superior fit to Models 1 and 2 (p<.0001, p=.0009, Wilcoxon signed-rank test).

H) Mean PRT per patch type, per subject from Model 3 simulations versus empirical mouse PRT, log-scaled (r^2 = .985), points colored per mouse.

I) Schematic depicted a sample trial for calculating model predicted PRT on single trials. Black trace indicates fraction of trials the model predicts subjects would still be on the patch at the start of that one second time bin (Top). Red dots indicate the probability of leaving, P(Leave), per that time bin. The product of fraction of trials remaining on patch times and P(Leave) determines the corresponding leave density for that time bin (Bottom). The average predicted PRT is then used as that trial's prediction.

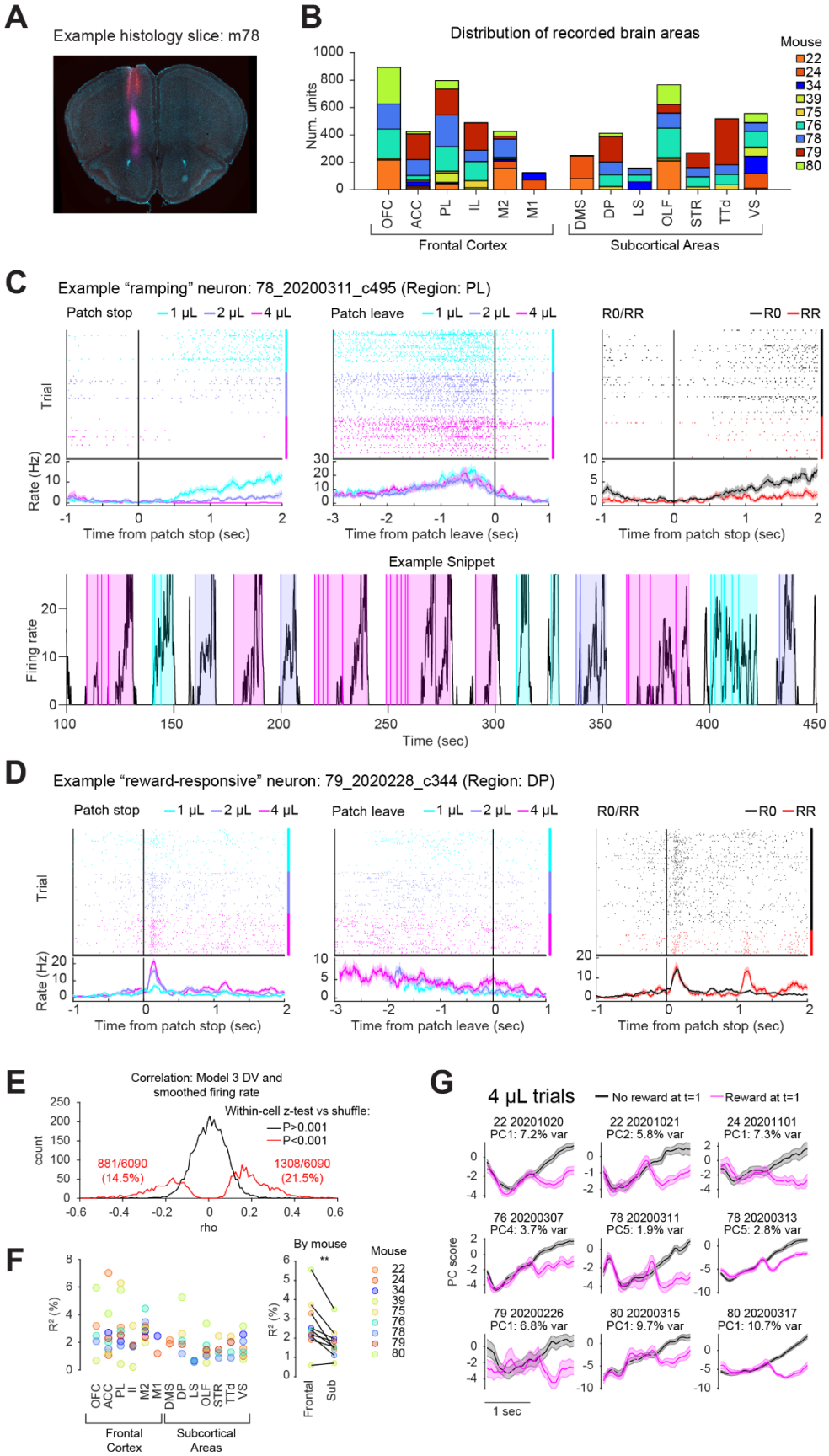
J) R^2 statistics for single trial predictions from cross-validated Model 3 fits across mice (Left, median r^2 = .544). Box edges indicate 25th and 75th percentiles. Whiskers stretch out to most extreme points, as none were considered outliers. Smoothed heatmap of single trial Model 3 predicted PRT versus empirical mouse PRT, sample mouse (Right). Predicted PRT for each trial was calculated using model fit parameters from training folds and compared with PRT over trials from held out test folds.

K) Model 3 predicted PRT and empirical mouse PRT across patches from the example session from Fig. 2F. Colored dots indicate mouse PRT, colored by reward size. Black dots indicate Model 3 predicted PRT, calculated from training folds. Gray lines show the difference in predicted versus

empirical PRT per trial. Red trace indicates latent estimations of patience for each patch. Y-axis is log-scaled.

L) Example schematics demonstrating how Models 2 (Left) and 3 (Right) make differing predictions for DV (Top) and P(Leave) predictions (Bottom) on patches with rewards at t=[0,2]s ('R0R' patches, black) compared with patches with rewards at t=[0,1,2]s ('RRR' patches, blue).

M) Per subject mean Simulated PRT for R0R versus RRR patches from Model 2 fits (Left), Model 3 fits (Middle), and empirical mice PRT (Right). Points are colored per mouse. Axes are log-scaled.



708

Figure 4: Introduction to neural recordings.

A) Example histology slice showing probe tracks targeting M2 and OFC. Red: DiI, Purple: DiD. Probe tracks were tilted relative to brain slices, so only part of each probe track is visible in each slice.

B) Distribution of recorded brain areas from 9 mice. Areas designated as "Frontal Cortex" are grouped together.

C) **Top:** An example neuron showing ramping activity whose slope depended on reward size and which was suppressed by reward delivery. Left: PSTH aligned to patch stop, split by reward size (cyan: 1 μL, purple: 2 μL, magenta: 4 μL). Middle: PSTH aligned to patch leave, split by reward size. Right: PSTH aligned to patch stop, split by whether reward was delivered at 1 second (red) or not (black); rewards of different size combined. **Bottom:** A snippet of the same neuron's activity across several consecutive trials, showing single neuron ramping activity over several seconds in individual trials. Shaded regions indicate patches, with the color indicating reward size. Vertical lines indicate reward delivery.

D) An example neuron showing transient activation to reward delivery. PSTHs as in C.

E) Histogram of Pearson's correlation between smoothed firing rate and the Model 3 decision variable (Figure 3) for each neuron in our data set. A shuffle distribution was generated separately for each neuron and a z-test was used to identify neurons with significant correlations.

F) R^2 values for the regression of Model 3 decision variable on individual neurons' firing rates (as in E), by brain region. In the left plot, each point represents a recording session. In the right plot, each point represents a mouse. Consistently across mice, frontal cortex areas had higher R^2 values than sub-cortical areas.

G) Hand-picked principal components of neural activity showing reward integrator-like activity. **Frontal Cortex Areas:** OFC: Orbitofrontal cortex, ACC: Anterior cingulate cortex, PL: Prelimbic cortex, IL: Infralimbic cortex, M2: Secondary motor cortex, M1: Primary motor cortex. **Other Areas:** DMS: Dorsomedial striatum, DP: Dorsal peduncular area, LS: Lateral septum, OLF: Olfactory areas, STR: Striatum, TTd: Taenia tecta dorsal part, VS: Ventral striatum. ** P < 0.01

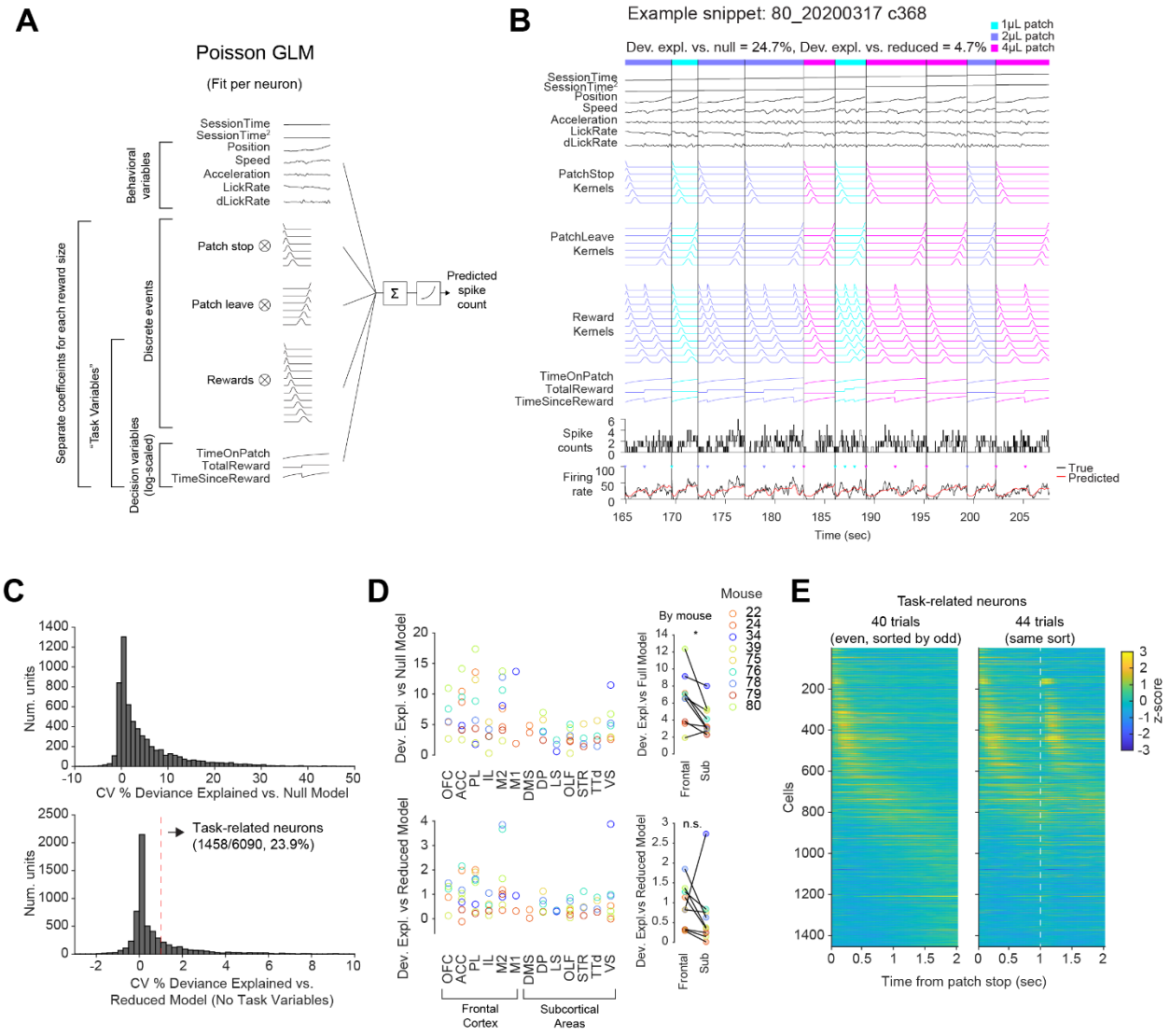


Figure 5: GLM modeling identifies subset of neurons with task-related activity.

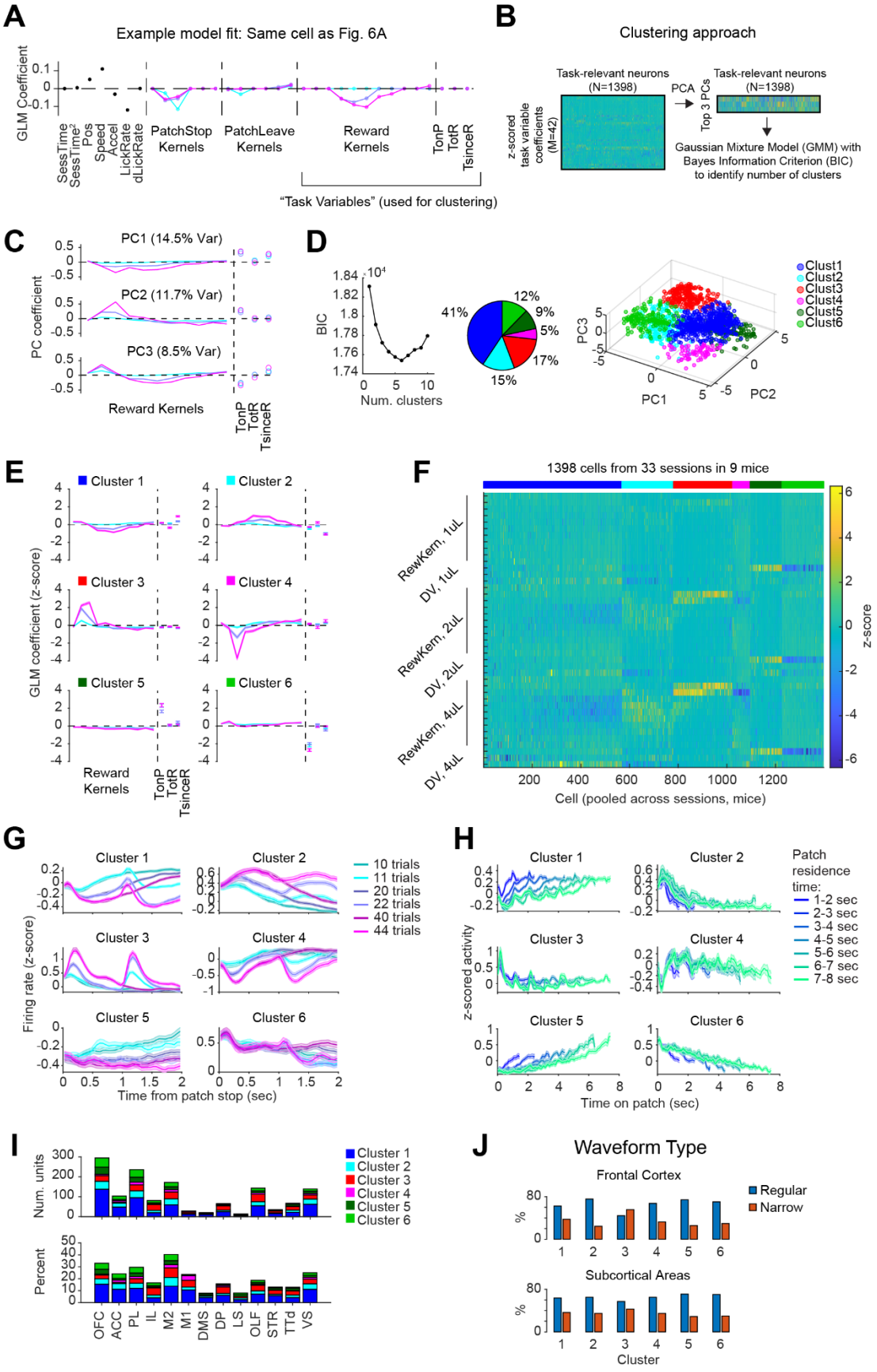
A) Schematic of the Poisson GLM.

B) A snippet of an example cell showing the structure of the Poisson GLM: Regressors (top) and neural activity (bottom). Inverted triangles in the bottom panel indicate reward delivery. Vertical lines indicate the boundaries between patches; only on-patch times were included. Colors indicate reward size. Reward Kernels, Time on Patch, Total Reward, and Time Since Reward were considered "Task Variables."

C) **Top:** Cross-validated percent deviance explained versus null model (intercept only) for all neurons in the dataset. **Bottom:** Cross-validated percent deviance explained versus reduced model (no task variables) for all neurons in the dataset. Neurons with >1% deviance explained versus reduced model were designated as "task-related."

D) Top: Average percent deviance explained versus null model by brain region. Within each brain region, neurons were averaged by mouse. Colors indicate mice. Bottom: Average percent deviance explained versus reduced model by brain region.

E) Z-scored neural activity for all task-related neurons from all brain regions on "40" trials (4 μL reward at 0 seconds, no reward at 1 second; left panel) or "44" trials (4 μL reward at 0 and 1 second; right panel; white dashed line indicates reward at 1 second). Neurons were sorted based on the time of peak activity on odd 40 trials; the left panel showed the same sort on even 40 trials, and the right panel shows the same sort on all 44 trials. An initial transient reward response (as in Figure 4D) gradually transitioned into upward ramping activity which was suppressed by reward delivery (as in Figure 4C).



762

Figure 6: Unsupervised clustering of GLM clustering reveals six clusters of neurons, the most prominent of which shows ramping activity and reward responses with opposite signs.

A) GLM coefficients for the example cell shown in Figure 5A. Task Variables were used for clustering (Reward Kernels, Time on Patch, Total Reward, and Time Since Reward; the latter three are also termed Decision Variables, or DVs). Only "Task-Relevant" neurons were included in clustering (Figure 5C), and further were required to have at least on non-zero Task Variable GLM coefficient.

B) Schematic of clustering approach. This approach was used to group task-relevant neurons into functional categories based on their response to task variables (reward kernels and decision variables).

C) The top 3 PCs of the matrix of Task Variable coefficients (accounting for 34.7% of variance). Task Variable coefficients were z-scored prior to PCA.

D) Gaussian mixture model (GMM) clustering on the top 3 PCs of Task Variables was used to identify clusters of neural activity patterns. Left panel: Bayesian information criterion (BIC) was used to select the number of clusters (minimum BIC: 6 clusters). Middle panel: Percentage of neurons assigned to each cluster. Clusters were ordered so that patterns with similar shapes but opposite signs were adjacent (see panel E). Right panel: Task-relevant neurons projected into the PC space used for clustering, colored by assigned cluster.

E) Average z-scored GLM Task Variable coefficients for each cluster.

F) Heatmap of z-scored GLM Task Variable coefficients for all neurons included in the analysis, split by cluster and sorted within cluster by the projection onto the first principal component (shown in panel B).

G) Average PSTHs of z-scored neural activity for each cluster, split by reward size and whether or not reward was delivered at 1 second (e.g. 10 indicates a 1 μL reward at 0 second, not 1 second, 11 indicates 1 μL reward at 0 and 1 second, etc).

H) Average PSTHs of z-scored neural activity for each cluster, split by patch residence time.

I) Top panel: Number of neurons in each cluster, split by brain region. Lower panel: Percent of all recorded neurons in each brain region assigned to each cluster.

J) Unit waveforms by cluster. Especially in Frontal Cortex, Cluster 3 was enriched for units with Narrow waveforms. 56% of Frontal Cortex Cluster 3 neurons were narrow spiking, versus 32% of Frontal Cortex neurons from the other 5 clusters. In Subcortical Areas, these numbers were 43% versus 34%.

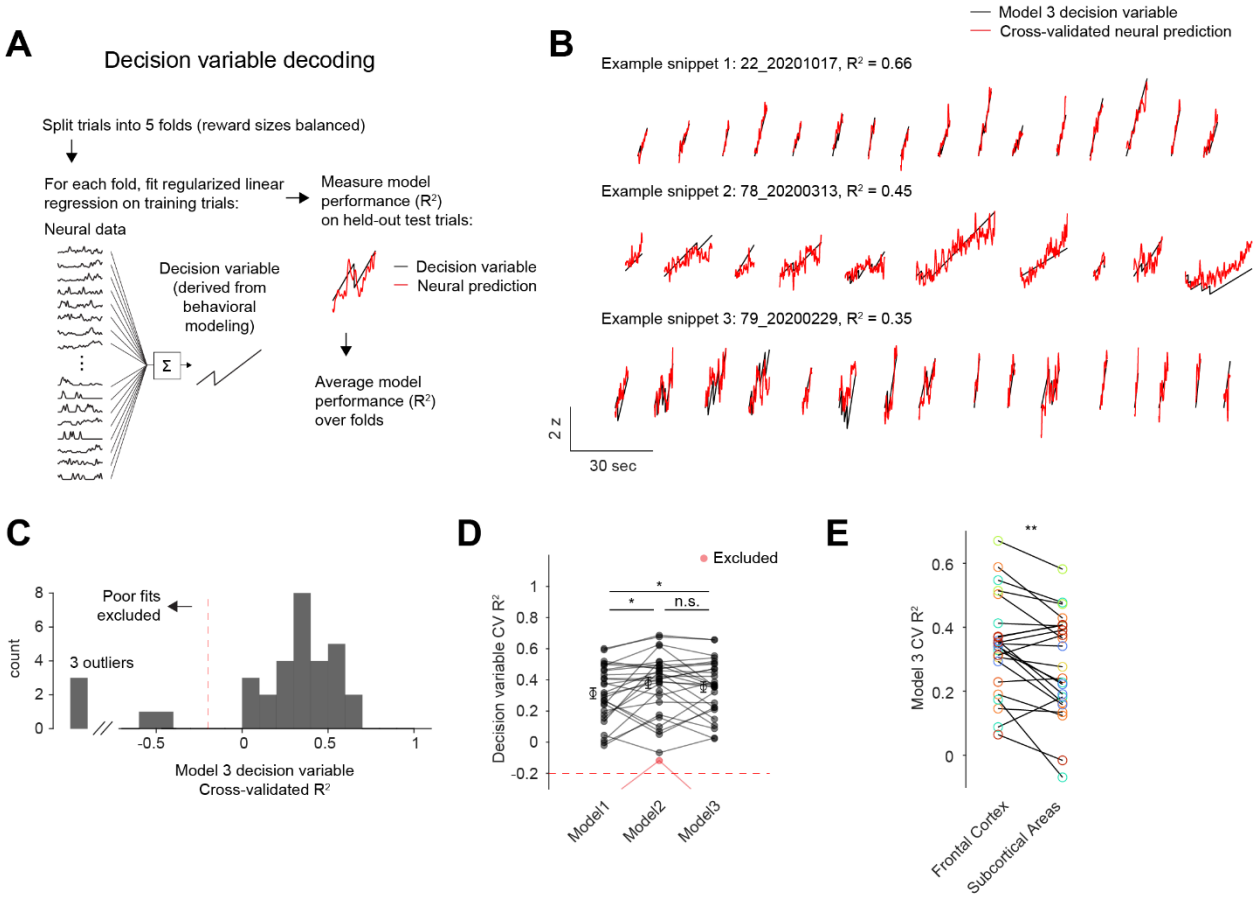


Figure 7: Decision variables derived from behavioral modeling are represented in neural dynamics.

A) Schematic of decision variable decoding (cross-validated linear regression of decision variables on smoothed, z-scored neural activity).

B) Snippets of several contiguous patches from three example recording sessions, showing the decision variable from Model 3 (Figure 3) in black and the cross-validated neural prediction in red. Neural predictions were generated using linear regression on training trials applied to held-out test trials.

C) Histogram of cross-validated R^2 for each recording session in the data set. Sessions with CV R^2 below an arbitrary threshold of -0.2 were deemed "poor fits" and excluded.

D) Comparison of CV R^2 for the decision variable derived from each of the models in Figure 3. * $P < 0.05$, Wilcoxon signed rank test (not corrected for multiple comparisons).

E) Comparison of Model 3 decision variable coding between Frontal Cortex and Subcortical Areas. For each within-session comparison, neurons were down-sampled so that each region had the same number of neurons. A minimum of 20 neurons per brain region per session were required. Only sessions with at least 20 neurons in each of the two brain regions were kept. Each pair of data points represents a recording session, and colors represent mice.

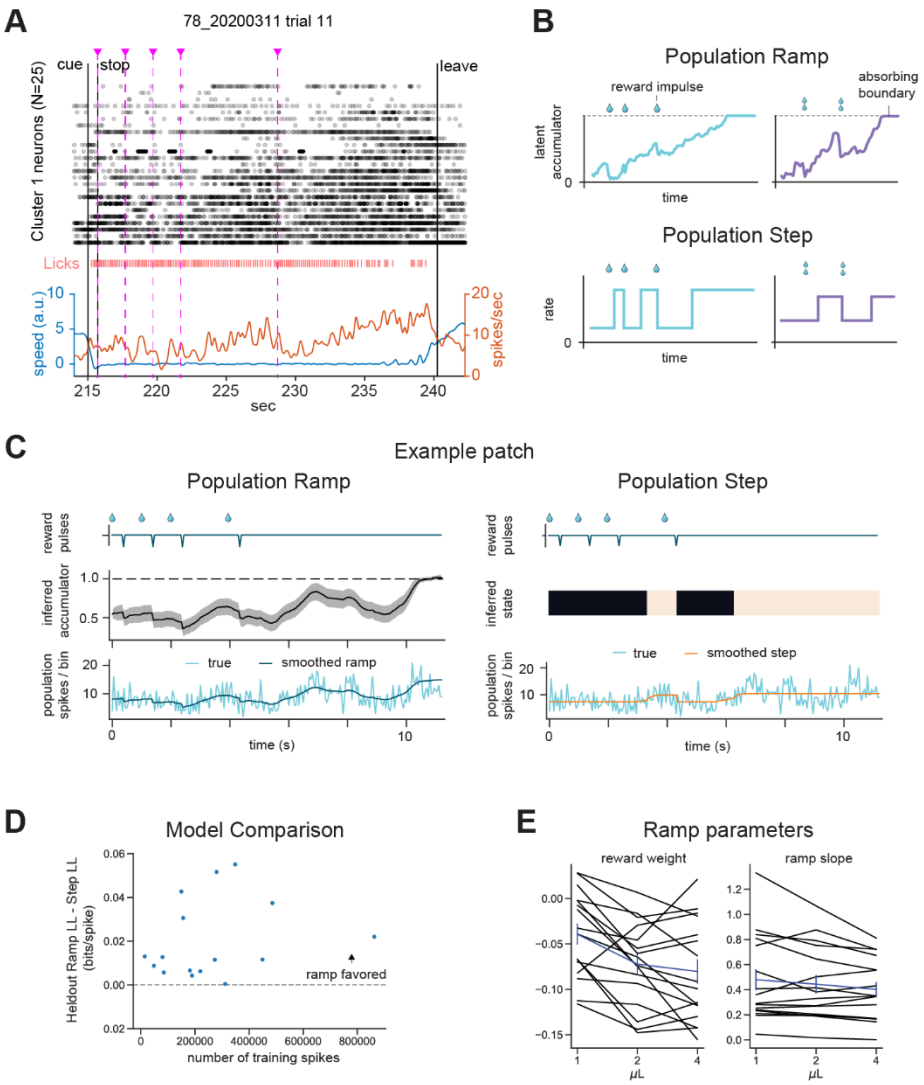


Figure 8: Population activity is ramping, not stepping, on single trials.

A) Example trial showing ramping activity of simultaneously recorded Cluster 1 neurons while the mouse remained still on the patch. Top: Raster plot of 25 Cluster 1 neurons. Middle: Raster plot of mouse licks. Bottom: Mouse speed (blue) and average firing rate of Cluster 1 neurons (red). Magenta dashed lines: Reward delivery (4 μL).

B) Schematic of ramp and step models. Models were fit to Cluster 1 neurons.

C) Single trial ramping and stepping model fits for an example session.

D) Model comparison across sessions. Only sessions with at least 10 Cluster 1 neurons were included (16/33 sessions). For 16/16 sessions, the ramp model outperformed the step model on held out data.

E) Ramping reward coefficients and slopes across reward size, by session.

Supplementary Figures

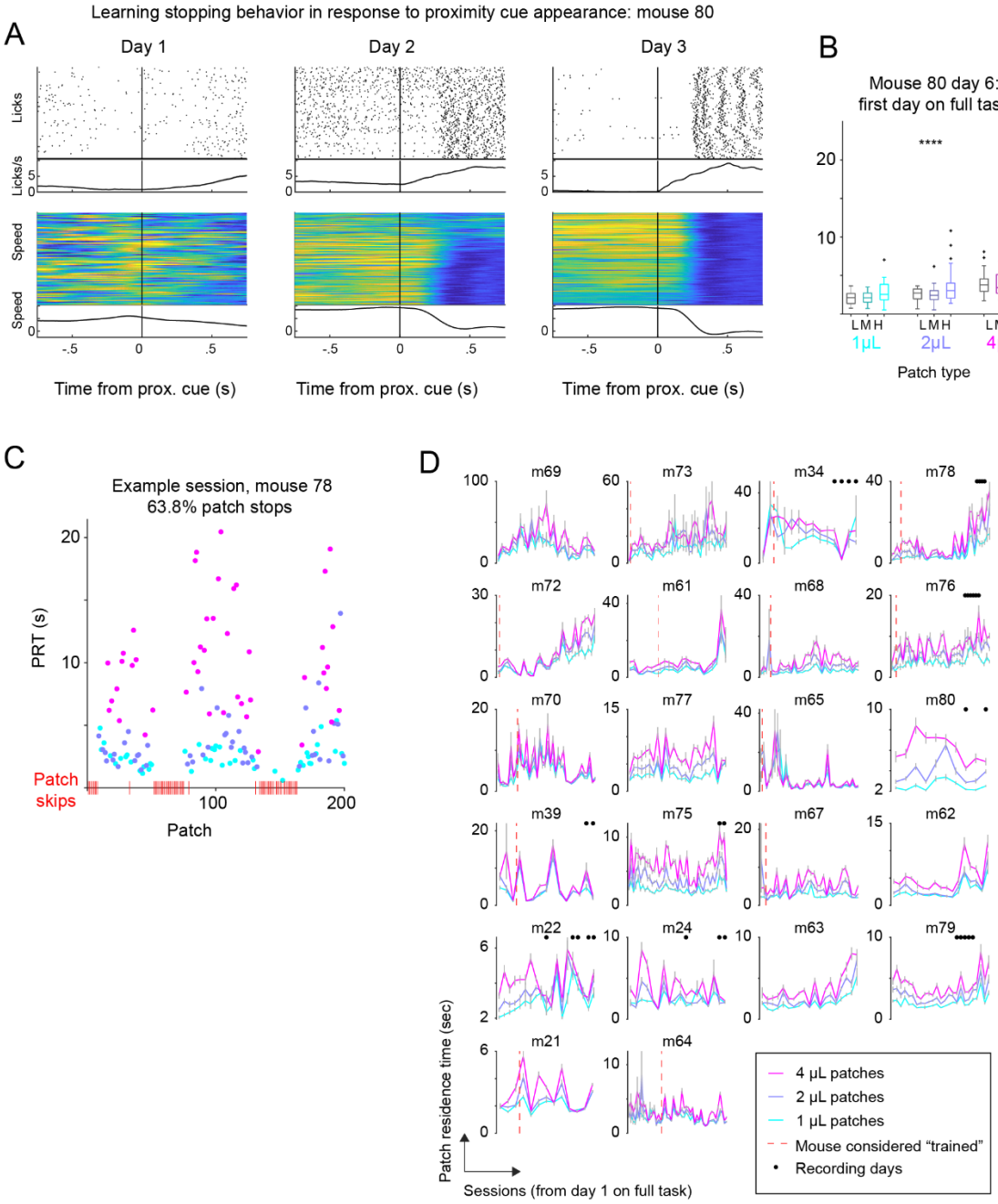


Figure S1: Mouse behavior over training (related to Figure 1).

A) Example mouse learning over first three days of pre-training in which proximity cue onset is followed by automatic reward delivery after a short delay. Patches are not introduced yet. Mouse learns that proximity cue is predictive of reward and begins showing anticipatory licking and slowdown in response to it.

B) First session full task, with patches introduced. Mice must stop in response to the proximity cue to enter a patch and earn rewards ($P < 0.0001$, linear regression of PRT on patch value). Column

groupings separate patches by reward size. Within columns, patches are split by reward frequency.

C) Example session in which mouse skips a substantial fraction of patches. PRTs are value-sensitive during periods of task engagement when mouse does stop to enter patches ($P < 0.0001$, linear regression of PRT on patch value). Colored dots indicate PRT on entered patches, colored by reward size. Red tick marks indicate patches skipped when mouse did not stop in response to the proximity cue.

D) Per subject PRTs across sessions, split by patch reward size. Error bars indicate standard error of the mean. Dashed red line indicates when mice task performance achieved inclusion criteria, after which sessions are included in the final data set. Mice with no dashed red line met inclusion criteria on the first day of the full task. Black dots mark recording sessions included for neural analysis.

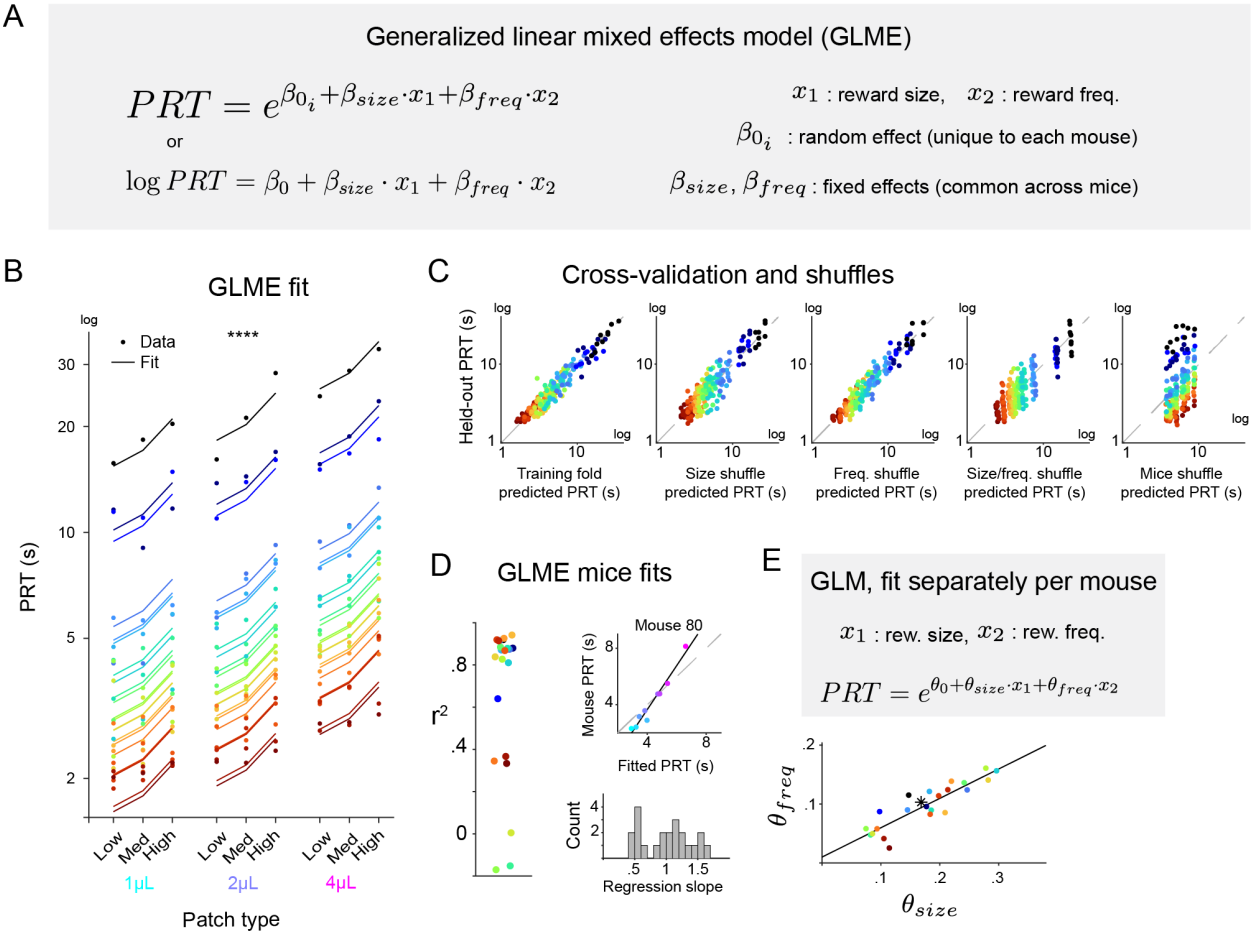


Figure S2: Generalized linear mixed-effects model captures behavior with fixed parameters across mice (related to Figure 2).

A) A generalized linear mixed-effects model (GLME) with log link function for predicting PRT across mice/conditions.

B) GLME-fitted and empirical PRT per mouse across patch types. Dots indicate empirical mouse PRT. Lines represent GLME fits. Colored by mouse ID.

C) Cross-validation of the GLME showing held-out PRT versus training fold predicted PRT. Training fold in the leftmost panel is unshuffled. Training fold trial labels for the remaining four were shuffled by reward size, frequency, size and frequency, and mouse ID, from left to right, respectively. Folds yielding the median mean squared error between predicted and held-out PRTs are shown for each shuffle type.

D) R^2 values using GLME fitted PRT as a predictor of empirical PRT per patch type (Left). Regression showing relationship between GLME fitted PRT versus empirical PRT, example mouse (Top, right). Slopes from per mouse regressions of empirical PRTs on GLME predictions (Bottom, right).

E) GLM predicting PRT as a function of reward size and frequency, fit separately per subject. Asterisk shows the fixed effects coefficients from the GLME fit across the population

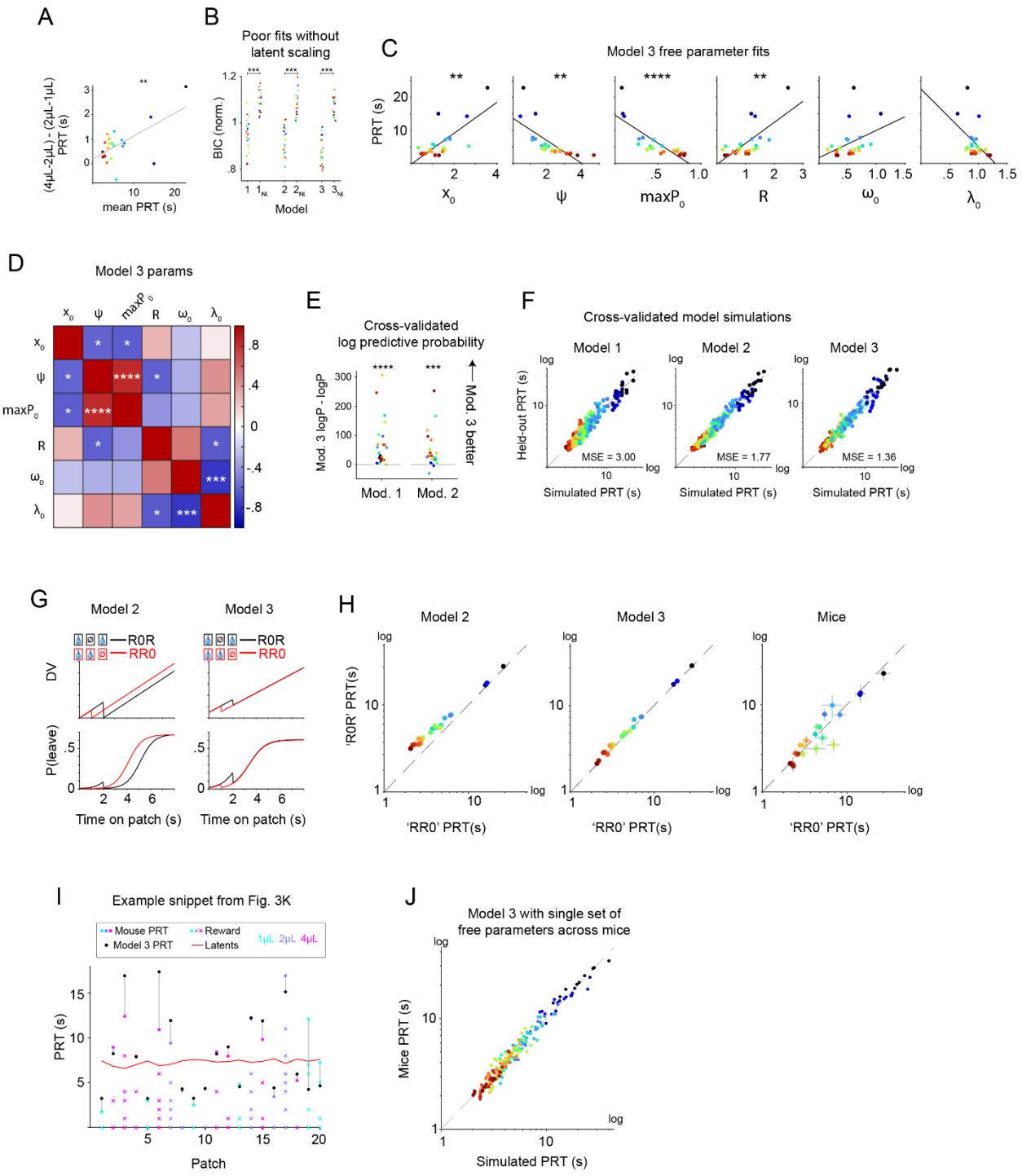


Figure S3: Additional analyses of behavioral model fitting (related to Figure 3).

A) Differences in PRT between 4μL-2μL patches versus 2μL-1μL patches scales with patience across mice (P = 0.0021, linear regression).

B) BIC values for model fits are consistently lower compared with non-latent-scaling counterparts across mice. 'NL' indicate no latent patience scaling (P = 0.0002 for each pairwise comparison, Wilcoxon signed-rank test, Bonferroni correction).

C) Model 3 free parameter fits versus mean PRT across mice (X_0: P = 0.009, Ψ: P = 0.0011, P_{max}: P < 0.0001, R: P = 0.0022, ω: P = 0.5828, λ: P = 0.2270, linear regression with Bonferroni correction).

D) Correlation matrix of Model 3 free parameter fit values across mice. Significant correlations indicated with asterisks, Bonferroni correction).

E) Relative log predicted probability of training fold model fits on held-out test folds, means across folds.

F) Mean model simulated PRTs across mice and patch types fit from training fold versus mean PRTs from held-out test folds.

G) Example schematics demonstrating how Models 2 (Left) and 3 (Right) make differing predictions for DV (Top) and P(Leave) predictions (Bottom) on patches with rewards at t=[0,2]s ('R0R' patches, black) compared with patches with rewards at t=[0,1,]s ('RR0' patches, blue).

H) Per subject mean Simulated PRT for R0R versus RR0 patches from Model 2 fits (Left), Model 3 fits (Middle), and empirical mice PRT (Right). Points are colored per mouse. Axes are log-scaled.

I) First 20 patches from Fig. 3K showing single trial mouse and Model 3 predicted PRTs with reward timings indicated per patch.

J) Simulated PRT versus empirical PRT for a Model 3 fit using a single free parameter fit across the population (MSE = 0.988).

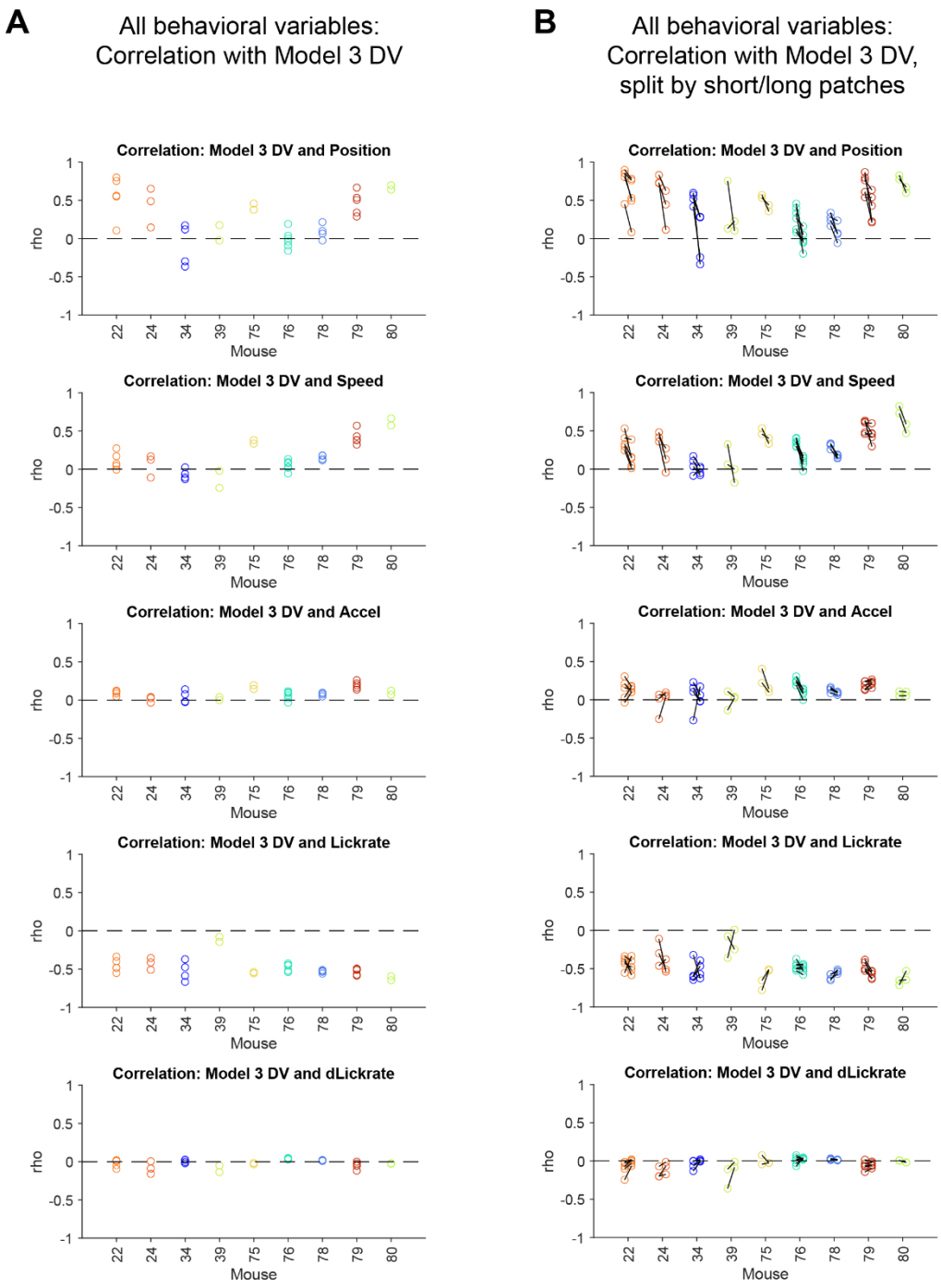


Figure S4: Correlations between behavioral variables and decision variables from behavioral model fitting (related to Figure 4).

A) Pearson's correlation between Model 3 decision variable (Figure 3) and multiple behavioral variables: position, speed, acceleration, lick rate, derivative of lickrate. Each data point represents a single Neuropixels recording session.

B) As in A, but split into short versus long patches (patch residence time above or below the median, per session). For each mouse, short patches are plotted on the left, and long patches on the right. Data points from the same session are connected by black lines.

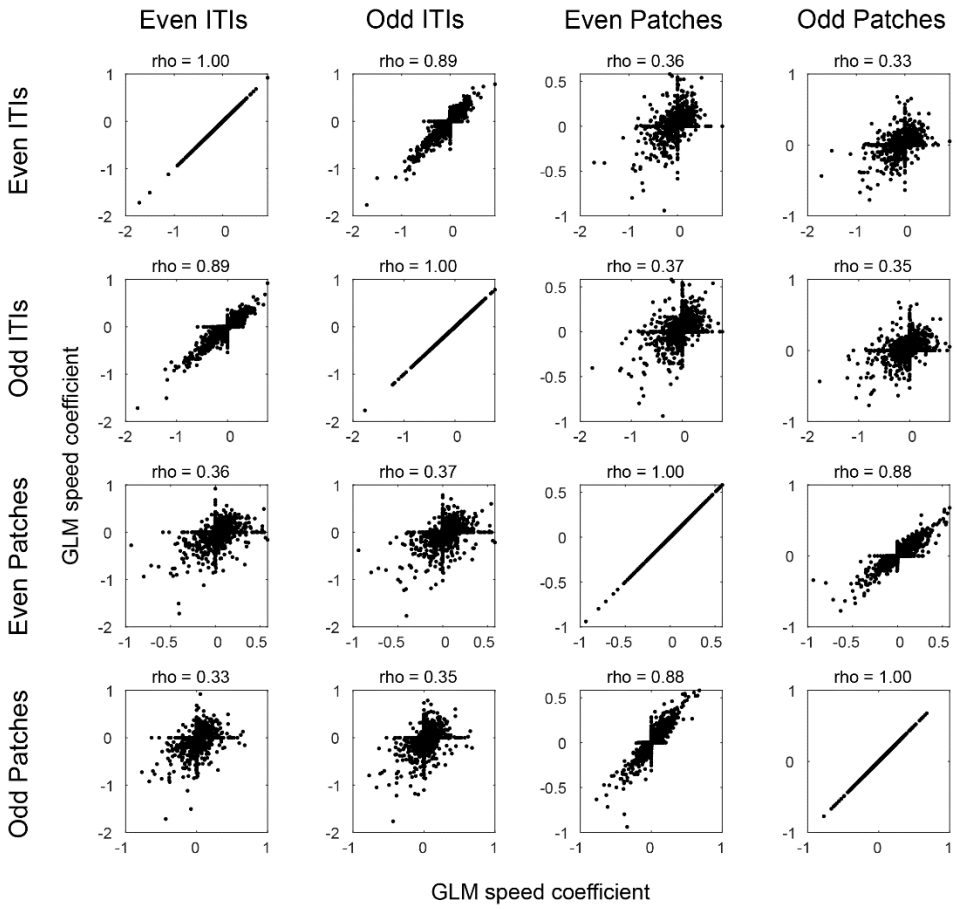


Figure S5: GLMs fit to intertrial intervals (ITIs) revealed qualitatively distinct speed coding compared to patches, suggesting that ramping activity in patches is not a low-level motor response (related to Figure 5). The structure of the GLM in patches was the same as in Figure 5, but fit to odd and even patches separately. The GLM in odd/even ITIs included the following regressors: Intercept, session time and its square, speed, and acceleration. Otherwise, the GLM fitting procedure in ITIs was identical to patches. In plots, each point represents a neuron, and "rho" above the plot indicates Pearson's correlation coefficient.

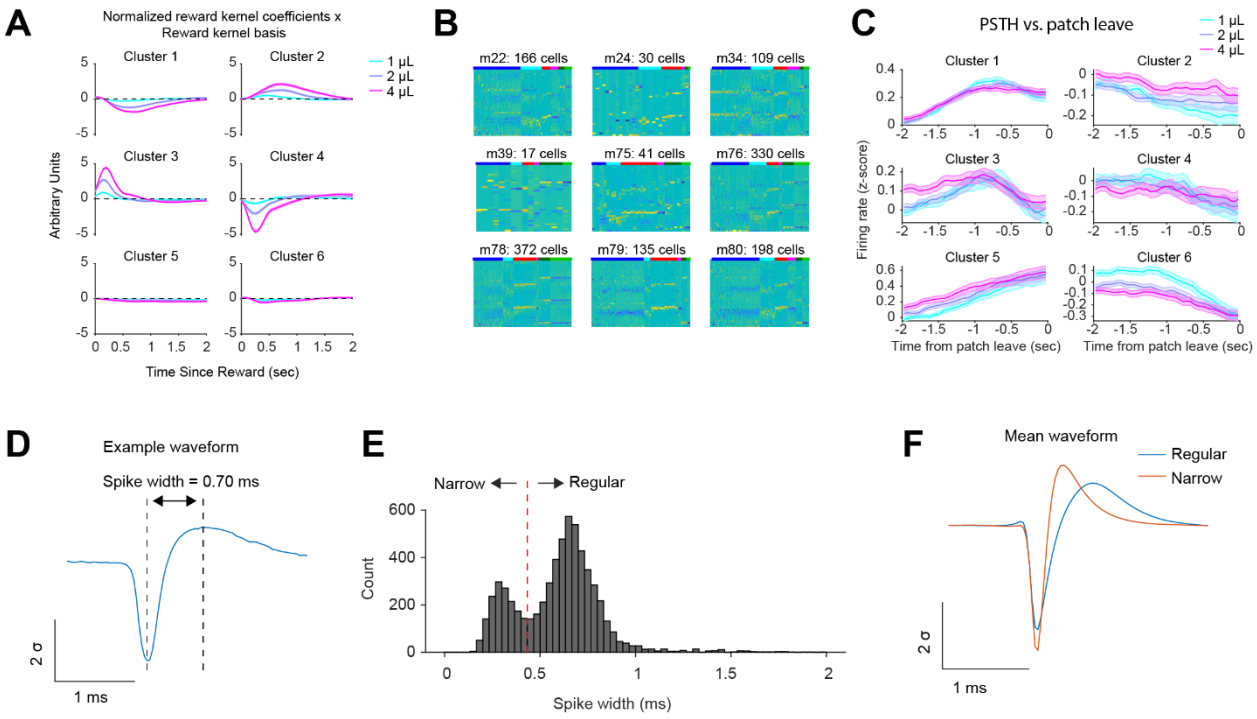


Figure S6: Additional analyses of GMM Clusters and spike waveform analysis (related to Figure 6).

A) The reward kernel coefficients from Figure 6 multiplied by the reward kernel basis to visualize the fitted reward response versus time for each GMM cluster.
B) Same as Figure 6F, but for each of the 9 mice in the dataset.
C) PSTHs of z-scored activity for each GMM Cluster, aligned to patch leave, split by reward size.
D) Example waveform indicating how spike width was measured.
E) Histogram of spike widths for all units in the dataset. A threshold of 0.433 ms was used to define Narrow versus Regular waveforms.
F) The mean waveform for Regular and Narrow units.

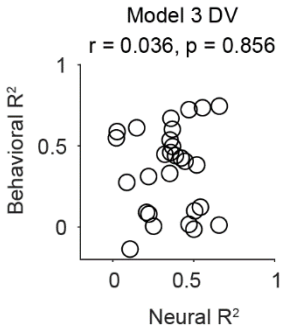


Figure S7: The ability of the Model 3 decision variable to predict behavior was not related to fidelity of neural decoding of that variable on a per-session basis (related to Figure 7). R^2 for Model 3 behavioral model fitting versus (y-axis) R^2 for decoding of Model 3 decision variable from neural data (x-axis), after excluding 5 sessions with Neural $R^2 < -0.2$.

Methods

Experiments

Mice

A total of 30 adult mice (C57/BL6j, 20 male, 10 female, age 2-5 months) were used in the experiments. All procedures were performed in accordance with the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Harvard Animal Care and Use Committee.

Surgery

Mice were anaesthetized using isoflurane (4% induction, 1-2% maintenance). For behavioral experiments, a custom titanium headplate was attached to the skull with Metabond (Parkell). For Neuropixels recordings, fiducial marks were made at the target sites for probe insertion using a fine-tipped pen, and a ground pin was inserted into the skull above contralateral cortex and attached with Metabond.

Virtual reality setup

Virtual reality setups were identical to those used in [45]. Three monitors (width 53 cm, height 30 cm) were placed in front and on either side of the animal. Virtual reality scenes were generated using VirMEn software in Matlab [46] on a workstation computer (DELL Precision 5810). Mice were placed on a cylindrical Styrofoam treadmill (diameter 20.3 cm, width 10.4 cm). The rotational velocity of the treadmill was recorded using a rotary encoder. The output pulses of the encoder were converted into

949 continuous velocity signal using custom Arduino code running on a microprocessor (Teensy 3.2). Velocity
950 was integrated within VirMEn to compute position. Water was delivered to the mouse from a spout
951 placed in front of the mouse's mouth. Licks were monitored using an infrared sensor (OPB819Z, TT
952 electronics). Voltage signals from the rotary encoder were digitized and recorded on the virtual reality
953 computer using a data acquisition system (PCIe-6323, National Instruments). Water timing and amount
954 was controlled using a solenoid valve (LHDA 1221111H, The Lee Company) and a switch (2N7000, On
955 Semiconductor), with TTL pulses generated by the virtual reality computer via the PCIe-6323 data
956 acquisition card.

957

958 *Neuropixels recording*

959 Spiking data were collected using Neuropixels 1.0 single shank probes. Neuropixels data were
960 recorded with SpikeGLX (https://billkarsh.github.io/SpikeGLX/). Craniotomies were performed the day
961 before recording and covered with Kwik-Cast (World Precision Instruments), and mice were allowed to
962 recover overnight. On the day of recording, a 3d-printed piece was placed over the head of the mouse,
963 blocking from the mouse's sight the experimenter manipulating the probe above the mouse's head. This
964 was critical to maintain good behavioral performance on Neuropixels recording days.

965 Neuropixels probes were lowered using a Thorlabs micromanipulator (PT1-Z8) at 9 μm/sec. After
966 reaching the target depth, the probe was allowed to settle for 30 minutes prior to starting the recording.
967 Recordings lasted 34 minutes (minimum) to 74 minutes (maximum) and were terminated once the
968 experimenter determined the mouse was no longer engaging with the task (e.g., running through
969 patches without stopping).

970

Spike sorting

Neuropixels data were spike sorted offline with Kilosort 2 or 3, followed by manual curation in Phy.

Histology

After the last Neuropixels recording session, mice were killed and perfused with phosphate buffered saline (PBS) followed by 4% paraformaldehyde in PBS. The brains were cut in 100 μm coronal sections using a vibratome (Leica). Brain sections were mounted on glass slides and stained with 4', 6-diamidino-2-phenylindole (DAPI, Vectashield). Slides were imaged with a fluorescence microscope (Zeiss Axio Scan.Z1). Probe tracks were identified in histology images and aligned to a reference atlas using SHARP-Track [47] or software written by Dr. Andrew Peters (https://github.com/petersaj/AP_histology).

Analysis

Unless otherwise noted, data were analyzed with custom code written in MATLAB.

Behavioral model fitting

Models 1 and 2 have five free parameters: X_0 sets the midpoint of the sigmoid. Ψ is inverse temperature. $\max P_0$ constrains the models' P(Leave) output. ω_0 is an exponent for power-law scaling of slope per reward size. λ_0 is an exponent for power-law scaling by latent state. Model 3 included those five, plus one additional free parameter for the value subtracted from the integrator per reward delivery, R.

Slope of the integrator model was scaled based on the patch's reward size. We fit this slope scaling by a power-law function to allow for varying sensitivity to reward size across mice. We used μ = [.5, 1, 2] for [1, 2, 4] μL, respectively, to normalize slope to the medium reward size.

$$\omega_{\mu} = \mu^{\omega_0}$$

Slope was additionally scaled by the estimate of patience for each patch (patch = i). We again used a power-law function for this scaling, fitting the free parameter, λ_0, per mouse to account for differences in how dependent PRT was on estimated patience across mice. Values for patience across mice were mean normalized to ensure fit values across parameters could be straightforwardly compared across mice. Each patch's normalized latent, L_i, was scaled by λ_0 to determine its contribution to slope scaling:

$$\lambda_i = L_i^{\lambda_0}$$

Because variance in patch leaving times increases when PRTs are longer, we also scaled maxP with patches' latent patience estimates. Because 0<maxP<1, we used a function that would reasonably allow scaling in either direction, while constraining the transformed value to remain between 0 and 1:

$$maxP_i = maxP_0/(\lambda_i * (1 - maxP_0) + maxP_0)$$

The other free parameters, Ψ, and, for Model 3, R, were not scaled across patches.

The models produce a decision variable (DV) per one second interval (t) by calculating its current integrator value, $X_{i,t}$ and subtracting X_0 from it:

$$DV = X_{i,t} - X_0$$

The models differ in how they each track time, TOP (Time on Patch) or TSLR (Time Since Last Reward), and whether they integrate rewards delivered (nRews = number of rewards delivered up to that point on the patch), to compute $X_{i,t}$:

Model 1, reward indifferent:

$$X_{i,t} = \frac{TOP}{\omega_\mu * \lambda_i}$$

Model 2, reward reset:

$$X_{i,t} = \frac{TSLR}{\omega_\mu * \lambda_i}$$

Model 3, reward integrator:

$$X_{i,t} = \frac{TOP}{\omega_\mu * \lambda_i} - nRews * R$$

To compute the model's output prediction, P(Leave), the probability of leaving a patch per one second bin, the DV is scaled by inverse temperature and sigmoid transformed with the output constrained by maxP:

$$P(Leave)_{i,t} = \frac{maxP_i}{1 + e^{-\psi DV_{i,t}}}$$

We optimized free parameter fits per subject and compared the models using mfit toolbox for Matlab. Maximum a posteriori estimates were computed using Matlab's fmincon function using 20 random initializations, converging upon the set of parameter fits that maximized log likelihood of the models at predicting P(Leave) across one second time bins. Ranges for the parameters used to fit were

X_0[-5:20], Ψ [0:10], maxP_0[.01:.98], ω_0 [0:2], λ_0[0:4], and R[0:20], with uniform priors for each. Resulting parameter fits are shown in Fig. S3. BIC values were computed for each model fit, and a protected exceedance probability was calculated for each model to estimate how prevalent the model is across subjects.

5-fold cross-validation was additionally performed across models. Trials were split into folds by ordering trials per subject and labeling them in serial 1-5. Fitting was then performed as above on the training folds, after which we computed the log predictive probability over trials from the held-out test folds.

Unit inclusion criteria

To be included for analysis, units from Neuropixels recordings had to have a minimum firing rate of 1 Hz within patches and additionally at least 1 Hz within each third of the session to ensure that they were tracked throughout the recording session.

Spike waveforms

Spike waveforms were extracted for each unit using C_Waves (https://billkarsh.github.io/SpikeGLX/). Spike width was defined as the difference between the time of the waveform minimum and the maximum following this minimum. A threshold of 0.433 ms was used to define Regular (>0.433 ms) versus Narrow (<0.433 ms) waveforms (Fig. S6D-F).

Generalized linear model (GLM) to predict neural activity from behavioral variables

Poisson GLMs were fit to spiking activity of single neurons using the glmnet toolbox (https://glmnet.stanford.edu). To prepare data for GLM fitting, we binned spikes from each neuron into 50 ms bins aligned to patch stop and continuing until patch leave. The following variables were included as regressors in the GLM:

1. Session time: time in session, (time in session)^2
2. Behavioral variables: patch position, velocity, acceleration, lick rate, d(lick rate)
3. Patch stop kernels: 6 kernels x 3 (1 set per reward size), peaks spanning 0-1 seconds from patch stop
4. Patch leave kernels: 6 kernels x 3 (1 set per reward size), peaks spanning 0-1 second prior to patch leave
5. Reward kernels: 11 kernels x 3 (1 set per reward size), peaks spanning 0-2 seconds after each reward
6. Time on patch: 3 (1 per reward size). Time on patch was log scaled: $t_{scaled} = \ln(t+1)$
7. Time since last reward: 3 (1 per reward size). Time since reward was log scaled: $t_{scaled} = \ln(t+1)$
8. Total reward: 3 (1 per reward size). Total reward was log scaled: $R_{scaled} = \ln(R+1)$

Thus, a total of 85 variables were used in the GLM (86 including intercept). Kernels were a raised cosine basis with log scaling of the x-axis (time), created using Matlab code from the Pillow lab (https://github.com/pillowlab/raisedCosineBasis). Parameters used were: logScaling = 'log', logOffset = 1.0, timeRange = [0 5].

Variables were z-scored prior to model fitting. Variables with one coefficient per reward size were z-scored across reward sizes, so that coefficients could be compared across reward sizes. Kernels were z-scored together.

Data were prepared for glmnet using custom MATLAB code, and then fit using glmnet in R. Model fitting was performed on the Harvard Faculty of Arts and Sciences Researching Computing cluster (FASRC cluster, https://www.rc.fas.harvard.edu/).

Models were regularized with elastic net regularization with $\alpha = 0.9$ (90% L1, 10% L2) to encourage sparse coefficients [48]. For cross-validation, patches were split into 5 folds. Each patch was assigned in its entirety to a single fold. Folds were counterbalanced for patches of different reward size. Cross-validation with these folds was used to select the regularization parameter for each neuron (using the function *cv.glmnet*).

Cross-validated deviance explained was computed for each neuron as follows:

$$D_{expl} = 1 - \frac{D_{model}}{D_{null}}$$

Where D_{model} is the cross-validated model deviance and D_{null} is the cross-validated null deviance. These two terms were computed once per test fold and then averaged across folds. For each test fold j, D^j_{model} and D^j_{null} were computed using the formula for Poisson deviance:

$$D^j_{model} = \frac{2}{N_j}\sum_i \left[y_i \ln \frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i) \right]$$

$$D^j_{null} = \frac{2}{N_j}\sum_i \left[y_i \ln \frac{y_i}{\mu_j} - (y_i - \mu_j) \right]$$

Where N_j is the number of observations (time bins) in the jth test fold, y_i is the number of spikes in the ith time bin in the jth test fold, $\hat{y}_i$ is the predicted mean of the Poisson distribution in the ith time bin (computed using the model fit on the training data), and μ_j is the mean spike count per bin in the training data. For bins with no spikes, we set $0 \cdot \ln(0) = 0$.

To estimate the contribution of task variables (reward kernels, time on patch, total reward, and time since reward) to model fits, we fit a reduced model in which task variables were removed. Poisson deviance was computed for the reduced model and deviance explained by task variables was computed as $1 - \frac{D_{full\ model}}{D_{reduced\ model}}$. Neurons with deviance explained > 1% by this metric were deemed "task-related."

Unsupervised clustering of GLM coefficients

We used a Gaussian Mixture Model (GMM) to cluster the matrix of Task Variables (M=42) x Task-Related Neurons (N=1398) (Fig. 6). The dimensionality of this matrix was reduced to 3 x 1398 using PCA. The dimensionality-reduced matrix was clustered using a Gaussian Mixture Model (function fitgmdist in MATLAB, with parameters RegularizationValue = 0.4, MaxIter = 10000, Replicates = 10, TolFun = 1e-6) with K (the number of clusters) spanning 1-10. For each value of K, the Bayes Information Criterion (BIC) was calculated. K was selected to minimize BIC (Fig. 6D).

Linear models to predict behavioral decision variables from neural activity

For each unit, spikes were binned into 50 ms bins and smoothed with a 100 ms kernel to generate single-cell firing rate traces within each patch. Then, firing rate traces were z-scored across patches for each neuron, so that model coefficients could be compared across neurons. We then fit cross-validated linear models to predict the following behavioral model variables from neural activity: the probability of leaving the patch in each time bin (p(Leave)), the ramping decision variable (DV), and for the relevant models, the total integrated reward (weightedRews). As described in the behavioral model fitting section, each of these was either scaled by the latent state (weighted average of surrounding patch residence times), or not, to test whether scaling by latent state improved model fits.

Regularized linear models were fit using the function *lasso* in Matlab with alpha = 0.1 (10% L1, 90% L2 regularization), predicting a given behavioral model variable from the activity of all recorded neurons at each time point. For cross-validation, patches were split into the same 5 folds as for the GLM (see above). The regularization parameter lambda was selected by nested cross validation within training

folds. Specifically, the training data was split into 10 folds and models were fit with a range of lambda values. Mean squared error was computed on nested test folds for each value of lambda, and lambda was chosen such that mean squared error was within 1 standard error of the minimum. The linear model was then fit on the entire training set to obtain β, the coefficients mapping neural activity to the behavioral variable, and R^2 was computed on the test set. R^2 and β were averaged across training folds to obtain final values for each recording session/neuron.

State space modeling of ramping and stepping dynamics

The ramping model was implemented as a constrained recurrent, switching linear dynamical system with Poisson emissions [49, 50]. The model has a continuous latent state x and discrete latent state z that both evolve over time. The discrete latent state can be in one of two states corresponding to an accumulation state ($z_t = acc$) and a boundary state ($z_t = b$). The continuous latent state is initialized to $x_0 < 1$. In the accumulation state, it follows a drift-diffusion process with a condition-dependent slope and diffusion variance σ^2. It also accumulates pulsatile reward inputs u_t. This corresponds to the following dynamics

$$x_t = x_{t-1} + \beta_{c,r} u_t + \beta_{c,s} + \epsilon_t, \quad \epsilon_t \sim Normal(0, \sigma_{acc}^2)$$

where $\beta_{c,r}$ is the coefficient for rewards in condition *c*, and $\beta_{c,s}$ is the slope for condition *c*. When $\beta_{c,r} < 0$ the reward decrements the accumulator, although importantly $\beta_{c,r}$ is unconstrained during fitting and is initialized at 0. In the boundary state, the continuous dynamics have no constant slope and a small constant variance. However, rewards still influence the continuous state which can cause the model to leave the boundary

$$x_t = x_{t-1} + \beta_{c,r} u_t + \epsilon_t, \quad \epsilon_t \sim Normal\left(0, \sigma_b^2\right)$$

where σ_b^2 is fixed to 1e-4. Transitions between discrete states depends on the value of the continuous state relative to a fixed boundary threshold at $x = 1$. In particular, the discrete state transition distribution is

$$p(z_t \mid z_{t-1}, x_t) \propto \exp\left(\gamma\left(\mathrm{R}_{z_{t-1}} + W x_{t-1}\right)\right)$$

where

$$R_{acc} = [0, -1]^T, \quad R_b = [0, -0.99]^T, \quad W = [0, 1]^T$$

and $\gamma = 50$. This setting of the parameters places a boundary at $x = 1$ with probabilistic transitions to and from the boundary state for values of the continuous state near the boundary. We note that the discrete state is slightly encouraged to stay in the boundary state, as in the discrete state the threshold for returning to the accumulation state decreases to $x = 0.99$. Overall, the generative model is

$$z_t \mid z_{t-1}, x_t \sim p(z_t \mid z_{t-1}, x_t)$$

$$x\_t \mid z_t, u_t \sim \begin{cases} Normal\left(x_{t-1} + \beta_{c,r} u_t + \beta_{c,s},\ \sigma_{acc}^2\right) & if\ z_t = z_{acc} \\ Normal\left(x_{t-1} + \beta_{c,r} u_t,\ \sigma_b^2\right) & if\ z_t = z_b \end{cases}$$

$$\lambda_t = \log(1 + \exp(C\, x_t))$$

$$y_t \sim Poisson(\lambda_t)$$

The time-varying reward inputs are shifted in time such that u_t is equal to one 400ms after a reward and otherwise equal zero.

The parameters $\theta = (\beta_{(1:C,r)}, \beta_{(1:C,s)}, \sigma_{acc}^2, C)$ were fit using variational Laplace EM (vLEM) [50], while the other parameters were fixed to correspond to the ramping model structure. The vLEM algorithm returns an approximate posterior distribution $q(x_{1:t})$ that approximates the true posterior distribution over the continuous latent states $p(x_{1:T} \mid y_{1:T}, u_{1:T}, \theta)$. For model comparison, the heldout log-likelihood was estimated via importance sampling using $q(x_{1:t})$

$$\log p_\theta(y_{1:T} \mid u_{1:T}) \approx \log \frac{1}{S} \sum_{s=1}^{S} \frac{p_\theta\left(y_{1:T}, x_{1:T}^{(s)} \mid u_{1:T}\right)}{q(x_{1:T}^{(s)})}, \quad x_{1:T}^{s} \sim q(x_{1:T}^{(s)})$$

with S = 200 and where the discrete latent state is marginalized in the numerator when computing p_θ.

The stepping model was implemented as a constrained HMM with input-dependent transitions and trial-dependent rates. It has two discrete states corresponding to a "down" and "up" state. Here the time-varying inputs u_t are one-dimensional and equal to 1 when there is a reward at time *t* and are otherwise zero. As in the ramping model, the rewards are shifted 400ms in time. The model starts in the down state and in the absence of a reward transitions to the up state with probability p_s. When there is a reward, the model is forced to transition back to the down state. The generative model is

$$\pi(u\_t) = \begin{cases} \begin{bmatrix} 1 - p_s & p_s \\ 0 & 1 \end{bmatrix} & if\ u_t = 0 \\ \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} & if\ u_t = 1 \end{cases}$$

$$z_t \mid z_{t-1}, u_t \sim \pi(u_t)_{z_{t-1}}$$

$$y_t \mid z_t \sim Poisson(\lambda_{z_t,i})$$

where $\lambda_{z_t,i}$ is the firing rate for state z_t on trial *i*. Notably, we fit separate firing rate parameters for each state and trial to allow the model to account for variability in overall rates across trials.

The overall model parameters are $\theta = (\lambda_{:,1:I}, p_s)$ and are fit via maximum likelihood using the EM algorithm. The per-trial firing rates are initialized to the mean number of spike counts in the first and last 10 time bins of a trial for the down and up states, respectively. To evaluate the model on heldout trials, the step probability p_s was kept constant but we fit the per-trial firing rates $\lambda_{z_t,i}$ to the test trials. Given the per-trial firing rates on test trials, the heldout log-likelihood was computed as the log-likelihood of the test trials marginalizing out the discrete states. This is a strong baseline since it has

access to the heldout data for fitting the per-trial firing rates, whereas the ramping model does not per-trial parameters that are fit on the heldout data.

For both models, we fit the model to the summed spike count of cluster 1 neurons from the frontal cortex. We fit and compared the models for sessions with at least 10 such neurons. The models were fit to binned spike counts in 50ms bins. For each session, 80% of the trials were used for training the model parameters and the held-out loglikelihood is computed and reported on the remaining 20% of the trials. Both models were fit using code in the SSM (https://github.com/lindermanlab/ssm) and SSM-DM (https://github.com/davidzoltowski/ssmdm) code packages.

Acknowledgements

We thank members of the Uchida lab for valuable discussions and comments on the manuscript. This work was supported by grants from the NIH Brain Initiative (U19NS113201, to S.W.L. and N.U.), the Simons Foundation (SCGB 697092, to S.W.L. and N.U.), the Sloan Foundation (S.W.L.), the McKnight Foundation (S.W.L.), and NIH F32MH126505 (M.G.C.).

Author Contributions

M.B. and N.U. designed the foraging task. M.B. built the behavioral setups with assistance from H.K. M.B., M.G.C., L.K., and J.S. collected behavioral data. M.B. and M.S.T. performed behavioral modeling with guidance from J.D. M.B., M.G.C., and L.K. collected Neuropixels data. M.G.C. analyzed Neuropixels data with feedback from M.B. and N.U. and contributions from M.B. D.Z. performed the recurrent linear dynamical systems modeling under the supervision of S.W.L. M.B., M.G.C. and N.U. wrote the manuscript.

Competing Interests

The authors declare no competing interests.

Data and Code Availability

Data and code will be posted to online repositories upon publication.

References

1. Krebs, J.R., J.C. Ryan, and E.L. Charnov, *Hunting by expectation or optimal foraging? A study of patch use by chickadees.* Anim. Behav., 1974. **22**: p. 953-IN3.
2. Stephens, D.W., J.S. Brown, and R.C. Ydenberg, *Foraging: Behavior and Ecology*. 2007: University of Chicago Press.
3. Marques, J.C., et al., *Internal state dynamics shape brainwide activity and foraging behaviour.* Nature, 2020. **577**(7789): p. 239-243.
4. Kolling, N., et al., *Neural Mechanisms of Foraging.* Science, 2012. **336**(6077): p. 95.
5. Greene, J.S., et al., *Balancing selection shapes density-dependent foraging behaviour.* Nature, 2016. **539**(7628): p. 254-258.
6. Hayden, B.Y. and M.E. Walton, *Neuroscience of foraging.* Frontiers in Neuroscience, 2014. **8**: p. 81.
7. Charnov, E.L., *Optimal foraging, the marginal value theorem.* Theor. Popul. Biol., 1976. **9**(2): p. 129-136.
8. Tajima, S., J. Drugowitsch, and A. Pouget, *Optimal policy for value-based decision-making.* Nature Communications, 2016. **7**: p. 12400.
9. Davidson, J.D. and A. El Hady, *Foraging as an evidence accumulation process.* PLoS Comput. Biol., 2019. **15**(7): p. e1007060.
10. Navabpour, S., J.L. Kwapis, and T.J. Jarome, *A neuroscientist's guide to transgenic mice and other genetic tools.* Neurosci Biobehav Rev, 2020. **108**: p. 732-748.
11. Bari, B.A., et al., *Stable Representations of Decision Variables for Flexible Behavior.* Neuron, 2019. **103**(5): p. 922-933.e7.
12. Kane, G.A., et al., *Rat anterior cingulate cortex continuously signals decision variables in a patch foraging task.* Journal of Neuroscience, 2022. **42**(29): p. 5730-5744.
13. Vertechi, P., et al., *Inference-Based Decisions in a Hidden State Foraging Task: Differential Contributions of Prefrontal Cortical Areas.* Neuron, 2020. **106**(1): p. 166-176.e6.
14. Cazettes, F., et al., *A reservoir of foraging decision variables in the mouse brain.* Nature Neuroscience, 2023: p. 1-10.
15. Henze, D.A., et al., *Intracellular Features Predicted by Extracellular Recordings in the Hippocampus In Vivo.* Journal of Neurophysiology, 2000. **84**(1): p. 390-400.
16. Cardin, J.A., et al., *Driving fast-spiking cells induces gamma rhythm and controls sensory responses.* Nature, 2009. **459**(7247): p. 663-667.
17. Latimer, K.W., et al., *Single-trial spike trains in parietal cortex reveal discrete steps during decision-making.* Science, 2015. **349**(6244): p. 184-187.
18. Zoltowski, D.M., et al., *Discrete stepping and nonlinear ramping dynamics underlie spiking responses of LIP neurons during decision-making.* Neuron, 2019. **102**(6): p. 1249-1258. e10.
19. Flavell, S.W., et al., *Serotonin and the neuropeptide PDF initiate and extend opposing behavioral states in C. elegans.* Cell, 2013. **154**(5): p. 1023-1035.
20. Levy, D.J., A.C. Thavikulwat, and P.W. Glimcher, *State Dependent Valuation: The Effect of Deprivation on Risk Preferences.* PLOS ONE, 2013. **8**(1): p. e53978.
21. Kane, G.A., et al., *Increased locus coeruleus tonic activity causes disengagement from a patch-foraging task.* Cogn Affect Behav Neurosci, 2017. **17**(6): p. 1073-1083.
22. Wikenheiser, A.M., D.W. Stephens, and A.D. Redish, *Subjective costs drive overly patient foraging strategies in rats on an intertemporal foraging task.* Proc Natl Acad Sci U S A, 2013. **110**(20): p. 8308-13.

23. Constantino, S.M. and N.D. Daw, *Learning the opportunity cost of time in a patch-foraging task.* Cogn Affect Behav Neurosci, 2015. **15**(4): p. 837-53.
24. Kane, G.A., et al., *Rats exhibit similar biases in foraging and intertemporal choice tasks.* eLife, 2019. **8**: p. e48429.
25. Cazettes, F., et al., *A reservoir of foraging decision variables in the mouse brain.* Nature Neuroscience, 2023. **26**(5): p. 840-849.
26. Lak, A., et al., *Orbitofrontal cortex is required for optimal waiting based on decision confidence.* Neuron, 2014. **84**(1): p. 190-201.
27. Murakami, M., et al., *Distinct Sources of Deterministic and Stochastic Components of Action Timing Decisions in Rodent Frontal Cortex.* Neuron, 2017. **94**(4): p. 908-919.e7.
28. Kane, G.A., et al., *Rat Anterior Cingulate Cortex Continuously Signals Decision Variables in a Patch Foraging Task.* J Neurosci, 2022. **42**(29): p. 5730-5744.
29. Luo, L., E.M. Callaway, and K. Svoboda, *Genetic dissection of neural circuits.* Neuron, 2008. **57**(5): p. 634-60.
30. Yizhar, O., et al., *Neocortical excitation/inhibition balance in information processing and social dysfunction.* Nature, 2011. **477**(7363): p. 171-8.
31. Lottem, E., et al., *Activation of serotonin neurons promotes active persistence in a probabilistic foraging task.* Nat. Commun., 2018. **9**(1): p. 1000.
32. Fonseca, M.S., M. Murakami, and Z.F. Mainen, *Activation of dorsal raphe serotonergic neurons promotes waiting but is not reinforcing.* Curr Biol, 2015. **25**(3): p. 306-315.
33. Stringer, C., et al., *Spontaneous behaviors drive multidimensional, brainwide activity.* Science, 2019. **364**(6437): p. 255.
34. Musall, S., et al., *Single-trial neural dynamics are dominated by richly varied movements.* Nature Neuroscience, 2019. **22**(10): p. 1677-1686.
35. Varela, F.J., E. Thompson, and E. Rosch, *The embodied mind: Cognitive science and human experience*. The embodied mind: Cognitive science and human experience. 1991, Cambridge, MA, US: The MIT Press. xx, 308-xx, 308.
36. Roitman, J.D. and M.N. Shadlen, *Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task.* J Neurosci, 2002. **22**(21): p. 9475-89.
37. Yartsev, M.M., et al., *Causal contribution and dynamical encoding in the striatum during evidence accumulation.* Elife, 2018. **7**.
38. Bahl, A. and F. Engert, *Neural circuits for evidence accumulation and decision making in larval zebrafish.* Nat Neurosci, 2020. **23**(1): p. 94-102.
39. Gold, J.I. and M.N. Shadlen, *The Neural Basis of Decision Making.* Annual Review of Neuroscience, 2007. **30**(1): p. 535-574.
40. Steinemann, N.A., et al., *Direct observation of the neural computations underlying a single decision.* bioRxiv, 2022: p. 2022.05. 02.490321.
41. Vijayan, V., et al., *A rise-to-threshold process for a relative-value decision.* Nature, 2023. **619**(7970): p. 563-571.
42. Maimon, G. and J.A. Assad, *A cognitive signal for the proactive timing of action in macaque LIP.* Nat Neurosci, 2006. **9**(7): p. 948-55.
43. Murakami, M., et al., *Neural antecedents of self-initiated actions in secondary motor cortex.* Nat Neurosci, 2014. **17**(11): p. 1574-82.
44. Hamilos, A.E., et al., *Slowly evolving dopaminergic activity modulates the moment-to-moment probability of reward-related self-timed movements.* eLife, 2021. **10**: p. e62583.
45. Kim, H.R., et al., *A unified framework for dopamine signals across timescales.* Cell, 2020. **183**(6): p. 1600-1616. e25.

46. Aronov, D. and D.W. Tank, *Engagement of neural circuits underlying 2D spatial navigation in a rodent virtual reality system.* Neuron, 2014. **84**(2): p. 442-56.
47. Shamash, P., et al., *A tool for analyzing electrode tracks from slice histology.* bioRxiv, 2018: p. 447995.
48. Driscoll, L.N., et al., *Dynamic Reorganization of Neuronal Activity Patterns in Parietal Cortex.* Cell, 2017. **170**(5): p. 986-999.e16.
49. Linderman, S., et al. *Bayesian learning and inference in recurrent switching linear dynamical systems*. in *Artificial Intelligence and Statistics*. 2017. PMLR.
50. Zoltowski, D., J. Pillow, and S. Linderman. *A general recurrent state space framework for modeling neural dynamics during decision-making*. in *International Conference on Machine Learning*. 2020. PMLR.