Population coding of strategic variables during foraging in freely-moving macaques

Neda Shahidi[1], Paul Schrater[2], Tony Wright[1], Xaq Pitkow[3*], Valentin Dragoi[1*]

[1] Dept. of Neurobiology and Anatomy
McGovern Medical School
University of Texas-Houston
Houston, TX 77030

[2] Dept. of Neuroscience,
University of Minnesota
Minneapolis, Minnesota

[3] Dept. of Neuroscience
Baylor College of Medicine
Houston, TX 77030

*** denotes equal contribution and corresponding authors**

Animals forage within their environment to extract valuable resources at lowest cost. Previous studies have suggested that animals simply maximize the current flow of reward without predicting the future outcomes of their actions and that this recent reward rate is represented in various brain areas. To test this, we devised a foraging task in which the relevant reward dynamics were hidden from the animal, and wirelessly record population activity in dorsolateral prefrontal cortex (dlPFC) while monkeys forage freely in their environment. We discover that their brains indeed contain predictions of future rewards and plans of their next actions. By decoding the dynamic reward probability and the memory of recent outcomes from the dlPFC population response, we show that monkeys create an internal representation of reward dynamics. The decoded variables predicted animal's subsequent actions better than either the true experimental variables or the raw neural responses. Our results suggest that the relevant task variables and behavioral decisions are dynamically encoded in prefrontal cortex during the time course of foraging.

Survival in environments with sparse resources requires animals to predict future outcomes before executing a costly action such as migration or relocation (Smith, 1982). Foraging for food reward represents an ideal paradigm for understanding the neural basis of planning and decision-making strategies in uncertain environments. Existing theories of foraging strategy, based on decades of behavioral experiments in various species, revolve around the idea of melioration or the matching law (Herrnstein, Rachlin, & Laibson, 1997), which states that an animal dedicates time or effort to an option proportional to its value. The majority of foraging studies use variable interval (VI) reward schedules in which rewards become available after random intervals in a history-independent manner (Aldiss & Davison, 1985; Gibbon, Church, Fairhurst, & Kacelnik, 1988; Herrnstein et al., 1997; Iigaya et al., 2019; Schneider & Davison, 2005; Sugrue, Corrado, & Newsome, 2004). These studies estimate value using the local or global history of reward availability, which suggests that the valuable option is chosen intentionally. However, as we show here, even a strategy that is blind to the reward rate and the future chance of reward can generate approximate matching, similar to what was observed in various species.

Previous studies have shown that the cognitive abilities of animals may be underestimated when animals are restrained (Freedman, 2008; Tollin et al., 2019). Furthermore, the effect of physical restraining might be even more dramatic on food-seeking behavior because animals use special locations or head movements for foraging (Bracis, Gurarie, Moorter, & Goodwin, 2015; Knight, 2011). Unfortunately, foraging studies performed in naturalistic environments did not attempt to record electrical activity in the brain. On the other hand, neurophysiological studies of foraging relying on head-fixed electrophysiological experiments in a laboratory environment used synthetic stimuli presented in a trial structure within a limited time frame, hence limiting the experimental conditions that could reveal the strategy that animals employ to optimize reward. We overcame the major limitations of these previous foraging studies by performing wireless recordings of population activity (Fernandez-Leon et al., 2015; Yin et al., 2013) in the dorsolateral

prefrontal cortex (dlPFC) while monkeys freely explored their environment to forage for food reward. Our approach allowed animals to freely interact with the task and explore the full range of reward expectancies. We discovered that unrestrained animals do not follow a blind strategy, but estimate the chance of future reward and use this estimation to make choices for future actions.

RESULTS

Monkeys (*n*=2) were offered two concurrent reward sources on a variable interval schedule whereby, at each source, food pellets became available ~10–30 seconds following previous reward. After minimally training the monkeys to press a knob and collect food pellets from concurrent reward boxes, animals interacted freely with the task equipment without having to learn a complex trial structure or being forced to respond within a short time frame. The two reward boxes were placed 120 cm apart (Fig. 1A, left) to create a cost for switching between the two sources of reward. A multi-electrode Utah array was chronically implanted in the dorsolateral prefrontal cortex (dlPFC) and spiking data was collected using a light-weight, energy-efficient wireless device (Fig 1A, right). The rewards on both sides (box 1 and box 2) became available at exponentially distributed random times (*i.e.,* constant hazard rates) with possibly different means (Fig. 1B, top two rows). Once a reward became available, it stayed available until the animal pressed the knob, at which time the reward was delivered (Fig. 1B, top three rows). This renders the probability of reward availability an increasing function of time elapsed since last response on the same box and the scheduled reward rate (see methods for the calculation of this probability; Fig S1). The inter-response time was determined by the monkey, leading to a probability of reward availability on the current box at the time of the response, p_{rew}, that was distributed over the range 0-1 (Fig. 1B, rows 4 and 5). By tracking the temporal evolution of the reward availability, the animal can better allocate its choices to optimize reward while minimizing travel costs compared to a strategy of matching local reward rates which are inversely proportional to the *loss count*, i.e. the number of consecutively unrewarded responses (Fig 1B, row 6).

We recorded the spiking activity of dlPFC neurons continuously throughout the session (*n*=1405, Fig 1C top row), and analyzed the response-locked events, *i.e.*, firing rates within two seconds before or after each knob response (Figure 2D). This allowed us to relate spike rates to events by averaging across choice and reward conditions (Fig 2C, bottom row).

We analyzed 33 sessions from two monkeys to understand the neural underpinnings of decision-making during foraging. We first confirmed that p_{rew} was predictive of whether the next knob press was rewarded (Fig 2A, right), while *loss count L* (Fig 2A, middle) and the fraction of rewards on one box, *i.e.*, its relative value (Fig 2A, left; see Methods), were not. This implies that an agent simply tallying the recent string of failed attempts to get reward cannot predict the reward, even if the proportion of responses over box 1 and box 2 for this agent follows the matching law. To confirm this, we simulated an agent that chooses to switch to the other box after a threshold *loss count* sampled from a Gaussian distribution. The parameters of this distribution were estimated from the distribution of *loss counts* at when the animal switched, accumulated

across 33 sessions of two monkeys (Fig 2B). This strategy corresponds to a simple win:stay / lose:switch rule. Although this agent is blind to both the average reward rate and the probability of the next reward, it nearly follows the matching law (Fig 2C). The slight under-matching that we observed resembles the behavior of various species in previous studies, while raising the possibility that some species followed a simple win:stay / lose:switch strategy that is blind to the actual rule of the game which predicts the chance of the reward.

We then simulated a second agent making choices based on the probability of reward availability on both boxes. The agent switches to the other side when the probability of reward availability on the other box exceeds that of the current box by a fixed switching cost, and otherwise waits for the probability of reward availability to increase everywhere (Fig 2B). Unlike the first agent, this agent has complete information about the task. Nonetheless, we again observed nearly matching behavior, now with slight over-matching (Fig. 2C and Fig. S3). The conclusion from the simulation of the two strategies is that approximate matching may be observed following a blind strategy as well as a fully informed strategy. Therefore, the matching law does not reveal the underlying strategy or the animals' ability to grasp the hidden rule of the task.

To understand if animal's strategy employs the hidden rules of the task, we predicted, for each attempt to gather reward, whether the animal stayed on the same box or switched sides. We based the prediction on two task variables: p_{rew}, the *loss count*. In 98% of rewarded responses, the monkey chose to *stay* at the same box, so we do not analyze these choices further. But for unrewarded responses, the chance of switching increased with both p_{rew} and *loss count* (Fig. 2D). Trained nonlinear binary classifiers (using radial basis functions) predicted decisions better than chance from p_{rew} but not from *loss count*. Predictions were best when using both variables (Fig. 2E; The median of the prediction performance was 0.7% above the chance level when *loss count* was used with p=0.26 for Wilcoxon signed rank with multiple comparison correction, 3.6% when p_{rew} was used with p=0.016 and 7% when both were used p=0.0005). Taken together, these findings are consistent with monkeys combining both task variables to decide what to do next (Fig. 2F). In principle, p_{rew} allows them to predict upcoming reward, while *loss count* allows animals to track the local reward schedule, in agreement with previous studies (Aldiss & Davison, 1985; Gibbon et al., 1988; Herrnstein et al., 1997; Iigaya et al., 2019; Schneider & Davison, 2005; Sugrue et al., 2004). This suggests that animals were able to infer and make decisions using a hidden task variable (p_{rew}).

Since p_{rew} and *loss count* predicted behavior well, we hypothesized that these variables were represented across the population of dlPFC cells. We first examined the population activity for the representation of p_{rew} before the reward is delivered. Neurons in the prefrontal cortex have been shown to modulate motor actions and their value [(Chandrasekaran, Peixoto, Newsome, & Shenoy, 2017; Kennerley & Wallis, 2009). However, there were only weak correlations between neuronal responses and task-irrelevant variables, such as spatial position within the environment and locomotion (Pearson correlation for location *X*: r=0.17 (p<<10^{-3}), location *Y*: r=0.12 (p=0.03), and locomotion: r=0.08 (p<<10^{-3}), Fig. 3A–B; also see Milton,

Shahidi and Dragoi, *in review*). To focus on reward-related variables, we removed the projections of location and locomotion from the neural activity (Figure 3C). Control experiments showed that eye movements have only a minor influence on brain activity while the animal interacted with the box, whereas they had larger movements and greater influence on neural activity during locomotion. Eye movements were correlated with locomotion during that time, with correlation r=0.16 ($p<10^{-3}$) for the eye velocity and r=0.13, $p<10^{-3}$) for the fixation rate (Milton, Shahidi and Dragoi, *in review*). Therefore, although we did not record eye movements during this task, since we removed the locomotion cues, this should attenuate the impact of the eye movements on the neural responses as well.

To examine the neural representation of task-relevant variables, we measured the spike count in the 1 second interval before the behavioral response (the "pre-response" interval from –1.1 to –0.1 seconds) for 1405 neurons. This time interval was selected because the modulation of neural activity was expected to start around 500 ms before the movement (Chandrasekaran et al., 2017), while the hand movement started approximately 500 ms before the response was recorded, so we simply combined the two time intervals. We observed that the response-averaged activity of some neurons differs for responses with the top 20% of p_{rew} from the bottom 20% (Fig. 3D) indicating that these neurons represent p_{rew}, and hence predict the reward. The average firing rates of these neurons within the pre-response interval was correlated with the log p_{rew} (Fig. 3E). This effect was observed for 36% of neurons in the entire dataset (Fig. 3F).

We further asked whether the population of simultaneously recorded neurons is able to estimate p_{rew} before each behavioral response. To test this, we trained a linear regression model to estimate p_{rew} from the pre-response neural activity (Fig. 3E). Chance predictions were determined by cross-validation on another regression model trained on shuffled responses. Reward was predictable in 31 out of 33 sessions when at least 5 neurons were used (Fig. 3F; $p \ll 10^{-3}$, Wilcoxon signed-rank test).

Consistent with previous reports (Elliott, Friston, & Dolan, 2000; Sugrue et al., 2004) neurons in dlPFC encoded the history of reward, quantified as *loss count* (Fig. S4). We further examined whether the neurons encoding either p_{rew} or *loss count* can predict the reward, before the response and thus before the actual reward is revealed. Since the reward itself is stochastic, its value at each press is unpredictable, but its average is predictable from p_{rew}. We thus used binary classifiers for which the inputs were the pre-response activity of either the p_{rew}-encoding neurons (those correlated with p_{rew} with $p<0.05$ and not correlated with *loss count*; $p>0.05$) or the *loss count*-encoding neurons (using analogous selection criteria). We found that the p_{rew}-encoding neurons predict reward better than chance (signed rank $p \ll 10^{-3}$) although the *loss count*-encoding neurons predict reward worse than chance (signed rank p=0.28). Thus, despite the stochastic nature of reward, dlPFC neurons can predict it prior to its delivery. Since dlPFC neurons encode both p_{rew} and *loss count*, and these two variables predict actions, we hypothesize that the neural representation of these two variables inform the animal's mental model of the task and drive its actions.

We next aimed to find neural representations of these task variables. Since the task variables and the neural activities were both internally correlated, we used canonical component analysis (CCA) to identify an uncorrelated pair of neural components maximally correlated with the task variables, p_{rew} and *loss count.* (Fig. 4A). Both canonical components contained a contribution from each task variable (Fig. 4B, inset; Fig S5). The task variables were more correlated with the canonical components than with any individual neuron (Fig. 4B and C), indicating that the information was distributed across the population.

We further examined whether the information encoded by neuronal populations can be used to predict animal's decision, *i.e.*, respond or switch. For example, the monkey could have a belief about reward availability that differs from the true probability p_{rew} in the experiment. Being wrong will produce behavioral variability (Beck, Ma, Pitkow, Latham, & Pouget, 2012) that could not be explained by p_{rew}. If the wrong beliefs are nonetheless correlated with the true probability, then the variability in beliefs could still be partially captured by the canonical components derived from the true probability, and hence these components could predict future choices even better than using our task variables directly. On the other hand, if the canonical components provide a sufficiently poor estimate of the animal's beliefs, whether because we don't have enough data from the right neurons or because the true probability is very different from the CCA fit, then the true reward probability could predict actions better than the estimates derived from the canonical components. When we select the two canonical components from the neural responses, and then use those signals to predict choice, we obtained better predictions than those made from the task variables themselves. In principle, the information encoded in the neural representation could predict future choices better or worse than our task variables, depending on whether decisions are driven by the variability within the neural representation. (Fig. 4D, E). We emphasize that the canonical components were not selected on the basis of actions, but rather were based on task dynamics (true reward probability and the number of consecutive losses); and yet these components were still predictive about actions. This indicates that the variability in these decoded dimensions is directly related to the animal's understanding of the task and could drive its choices.

The variability manifested in the CCA estimates of the task variables is a subset of all neural variability. Some of the remaining variability could reflect other interpretable variables about the task that we have not considered, and some could reflect other sensory inputs, movements, thoughts, mental states or processes, or irrelevant variability ('noise'). Any of these sources of variability may predict an animal's choice. Since the neural activity contains the canonical components, we expect to be able to predict decisions at least as well from the full population as from those two components. However, many more parameters need to be estimated when linearly predicting choice from the full population, and this could lead to overfitting, and thus worse performance on held-out data (Fig S6). Indeed, we found that cross-validated predictions from the neural representation of the two task variables were better than predictions from the full population (Fig. 4D, E). These findings show that the task-targeted dimensionality reduction of neural responses provided a useful description of information the monkeys used to perform the foraging task.

DISCUSSION

We discovered that monkeys perform foraging decisions based on reward probabilities inferred indirectly from ambiguous evidence. Our neural population analysis reveals that these probabilities are encoded in the brain, and this encoded information predicts the animal's future actions. Remarkably, the encoded probabilities predict animal's choices even better than either the true reward probabilities or the entire population activity. This is an important demonstration of how targeted dimensionality reduction can reveal neural computations better than behavior or unprocessed neural activity.

By employing a freely moving approach, our study is a pioneering move toward studying neural correlates of cognition in a free-roaming setting. Previous studies have underestimated the cognitive capacity of monkeys during foraging, primarily because of their restrictive experimental paradigms. The free-roaming setting also enabled up to implement the switching cost between two reward options as simply allowing the monkey to walk between them. This is commonly implemented as a timeout period immediately after switching decisions, which potentially alters neural responses in dlPFC. It is also possible that the higher arousal state of the monkey in the free-roaming setting [Milton, Shahidi, and Dragoi, *in review*] has enhanced their cognitive ability to perform the task.

Taken together, our findings challenge long-standing theories of reward-seeking behavior, suggesting that animals maximize the recent rate of reward, without constructing a reward model to predict the future. We propose a novel paradigm eliminating a fixed trial structure, in which observing the animal's behavior is preferred to restraining it. This minimizes the interference with the decision-making process in the animal's brain (Calapai et al., 2017; Fagot, Gullstrand, Kemp, Defilles, & Mekaouche, 2014) and alleviates the need for training the animal for complex associations which could alter the population dynamics in associative areas of the brain such as the lateral PFC (Bunge, Burrows, & Wagner, 2004). This paradigm shift has been suggested decades ago (Hernandez-peon, Scherrer, & Jouvet, 1956), but is only feasible now due to advances in low-power, high-throughput electrophysiological devices as well as large scale computing. We believe this shift toward more natural behavior is inevitable for the future of neuroscience.

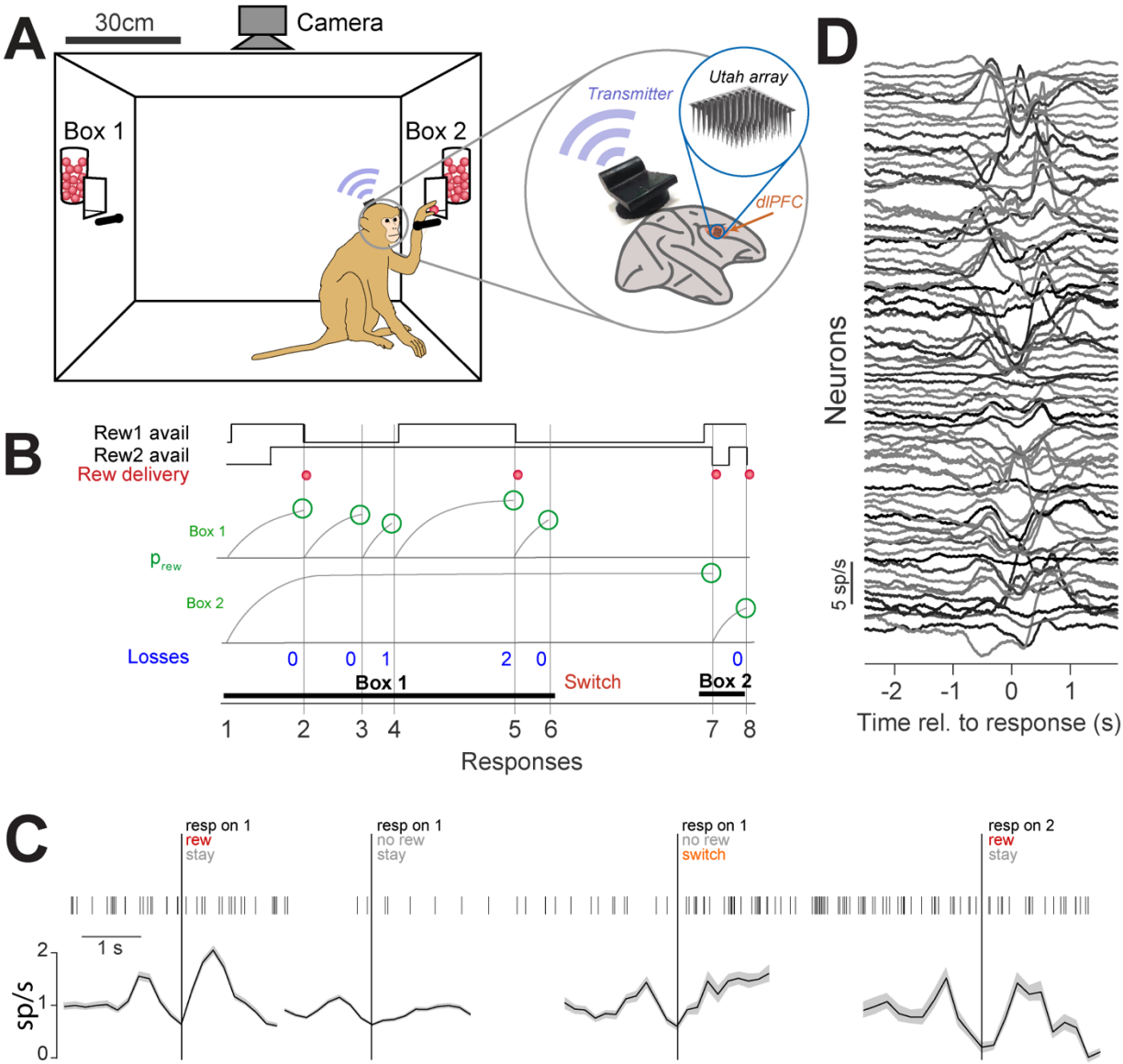

Figure 1 – Foraging in freely-moving monkeys while population activity in prefrontal cortex is recorded wirelessly. (A) *Left*: Schematic of the experimental cage with two reward boxes, two knobs, and an overhead camera. *Right*: Location of the Utah array in dlPFC (area 46) and wireless transmitter roughly scaled to the size of the schematic of the brain. **(B)** Illustration of task dynamics with 8 hypothetical responses in the concurrent variable interval foraging task. In this illustration, the monkey responds six times on box 1, then switches to box 2 and responds twice. Therefore, response 6 is considered a response with a *switch* choice. The first two rows show the independent Bernoulli processes determining when rewards became available at boxes 1 and 2. The mean values of the Bernoulli distributions are determined independently at the beginning of the experiment and switch between box 1 and 2 after 34 and 134 rewards were delivered. After 100 rewards, there was a one-hour break and the schedule may or may not switch after the break. Mean time intervals of the switching process for reward availability were selected uniformly from 10, 15, 20, 25 and 30 s. Reward availability was hidden from the monkey. If the reward was available at the time of the response, it was delivered, and then became unavailable. In the example shown, response

numbers 2, 5, 7, and 8 were rewarded (third row). p_{rew} (4th and 5th row) is shown at the time of each response (green circles; see Methods; Fig S1). The *loss count* (6th row) is defined as the number of consecutive unrewarded responses at the current box before the current response. **(C)** The spike-train of an example neuron on the timescale of four knob responses (top row). Spike trains were cut into response-locked intervals of identical length. The bottom row shows the firing rate of the same neuron, averaged across responses with reward/no reward and stay/switch. We chose a low-firing neuron for the clearest visualization of the spike raster. However, the average firing rates of the population of neurons were higher (See panel D). **(D)** Response-averaged firing rates of 80 single and multi-units recorded simultaneously.

Figure 2

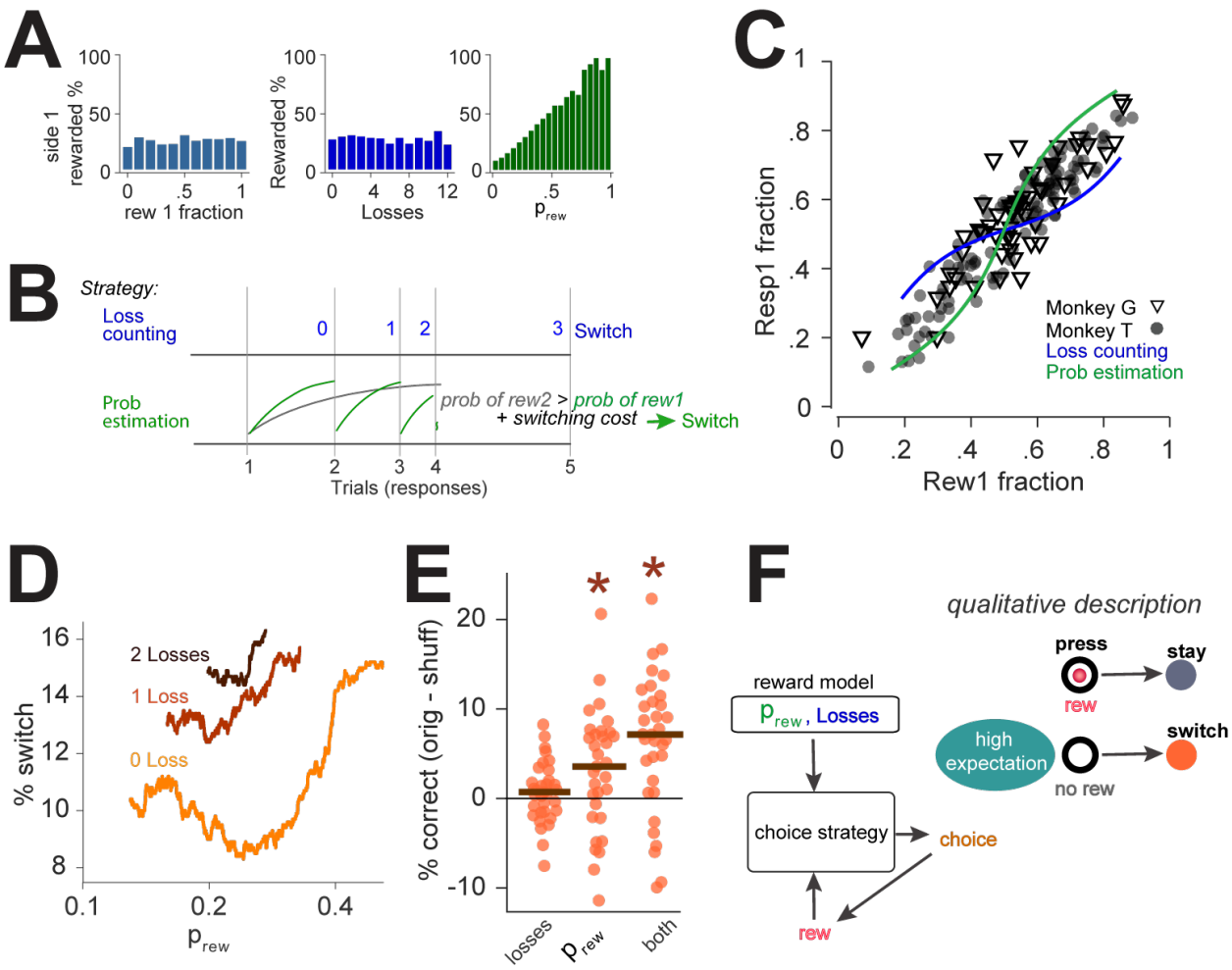

Figure 2 – Behavioral and neuronal performance during foraging. (A) Histogram of 13269 rewarded and unrewarded responses from 33 sessions, showing the dynamic reward ratio (Sugrue et al., 2004), the *loss count*, and p_{rew}. **(B)** Illustration of foraging strategies for two simulated agents. The '*Loss counting*' agent switches to the other box when *loss count* exceeds a threshold drawn from a Gaussian distribution. The 'probability estimator' agent switches to the other box when the probability of reward availability on the other box exceeds p_{rew} by a fixed switching cost (Charnov, 1976). The inter-response times were drawn from a random geometric distribution for both agents. The parameters of these strategies, namely the *loss count* distribution, the inter-response time distribution, and the switching time, were estimated from the behavior of the monkeys. Each agent was simulated for 100 rewards. The variable interval (VI) reward schedules spanned the range between VI-5 and VI-50 in steps of 1 s, and were drawn independently for each box. **(C)** Near-matching behavior of two monkeys and two simulated agents. For each set of schedules, response 1 fraction represents the number of responses in box 1 divided by the total number of responses as a function of rew 1 fraction (26 sets of schedules for monkey G and 59 sets of schedules for monkey T are shown). The same ratios were calculated for each set of schedules for simulated agents in (B). **(E)** The percentage of *switches* as a function of p_{rew} and *loss count*. Responses with 0, 1, and 2 losses were accumulated across all sessions, sorted according to p_{rew} and smoothed with a window of 1000 responses. Responses with *loss count* > 2 were ignored for this analysis since they had < 1000 responses. **(E)** The prediction of choice (stay/switch) using a support vector machine with RBF kernels, trained and cross-validated for each session (n=33). The predictors were the logarithm of *loss count*, the logarithm of

p_{rew}, or both. We used the unrewarded responses only because the fraction of responses in which monkeys switch after a reward was very low. The number of *stay* and *switch* training observations were equalized by bootstrapping. This analysis was repeated with shuffled observations, and then decoder performance with shuffled observations was subtracted from that of the original data. **(F)** A model of decision making in the foraging task that uses an internal model of reward availability and the actual reward to decide whether to make a choice. The internal model of the reward availability relies on an estimation of p_{rew} and *loss count*.

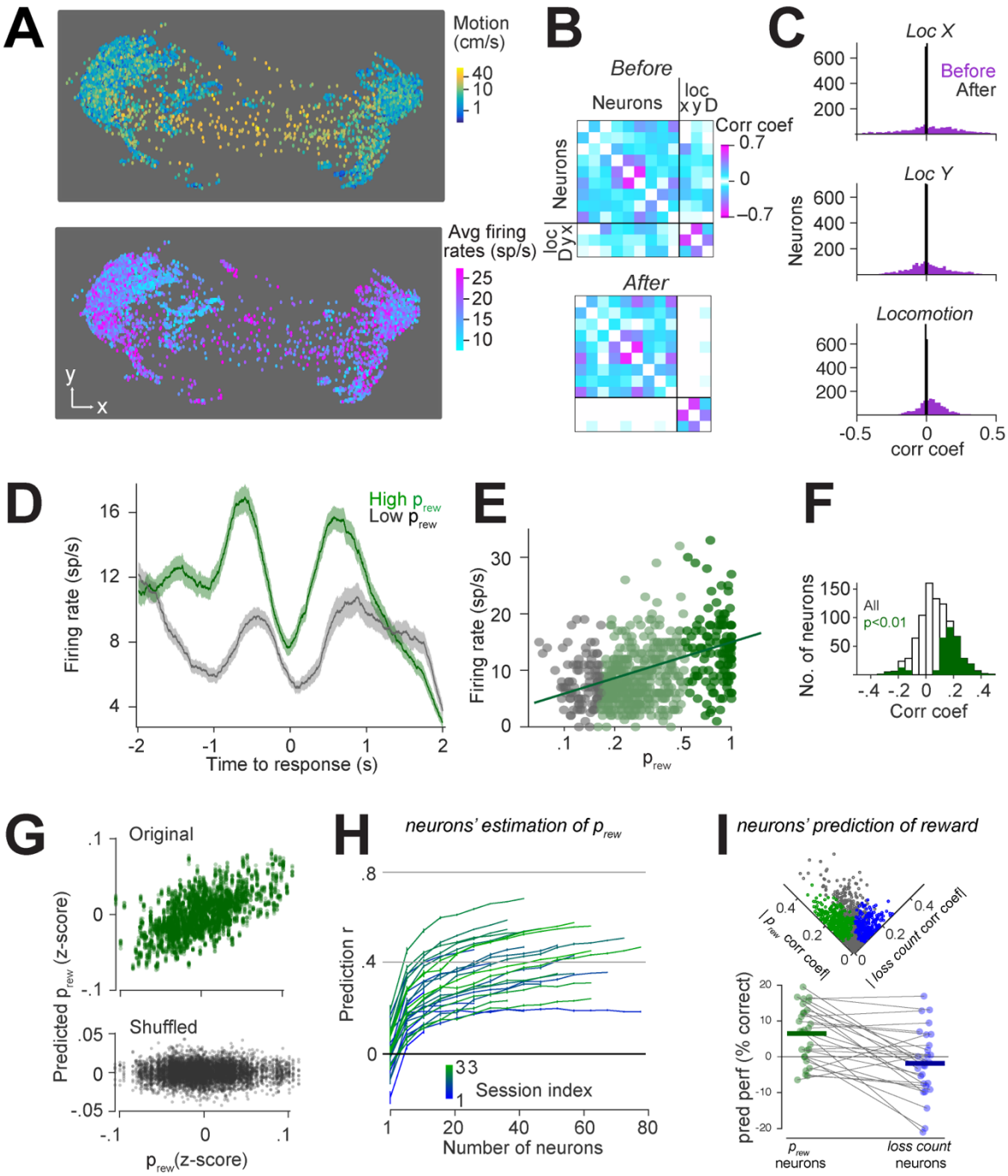

Figure 3 – Neuronal populations predict reward by estimating p_{rew}. (A) *Top*: The location and locomotion of the monkey in a sample session. Only time bins < 3 s before or < 1 s after a response were shown (each time-bin is 200 ms). *Bottom*: Population-averaged firing rates in a sample session for each time bin in (A). **(B)** Correlation coefficient between the task-irrelevant variables (location and locomotion of the monkey) and the pre-response firing rate of each neuron, before and after decorrelation by Gram Schmidt orthogonalization. This decorrelation was applied to the 3-dimensional space of location and

locomotion (speed). For illustration, an arbitrary subset of neurons is shown. **(C)** Correlation coefficients as in (B) but for all recorded neurons. **(D)** A sample neuron for which the pre-response firing rate modulates p_{rew}. The firing rate was calculated for each 200 ms time bin starting 2 s before and ending 1 s after the responses. Firing rates were averaged over responses with low (< 20th percentile) p_{rew} (gray) and high (> 80th percentile) p_{rew} (green). **(E)**: For the same neuron, the firing rate of the 1 s interval pre-response (-1.1 to -0.1 s) against p_{rew} on a logarithmic scale. The linear correlation coefficient between log p_{rew} and firing rate was 0.37 ($p<<10^{-3}$). **(F)** The pre-response activity of 36% of neurons was significantly correlated with p_{rew} *(p<0.01;* positive correlation: 29% of neurons; negative correlation: 7% of neurons). **(G)** Prediction of p_{rew} in all responses of a sample session, using the pre-response activity of simultaneously recorded neurons in this session. The predictor was a cross-validated regression model with RBF kernel and support vectors. The correlation between the predicted p_{rew} on the y-axis and actual p_{rew} on the x-axis was used to quantify the goodness of fit. For this session, correlation was 0.55 and for a model trained and tested on shuffled responses, the correlation was -0.017. **(H)** Prediction performance for 31 out of 33 sessions as a function of the number of neurons used as predictors. The predictor neurons were chosen randomly from the population. The random selection was done 50 times for each data point (starting with 1 neuron at a time, then increasing up to the number of neurons in the session with steps of 5 neurons). **(I)** Reward prediction using the pre-response activity in each session using either the neurons whose responses were correlated with p_{rew} ($p<0.05$), but not with *loss count* ($p>0.05$), or vice versa. The predictors were cross-validated support vector machines with radial basis function kernels. The performances of shuffle-trained predictors were subtracted from the performance of the original predictors.

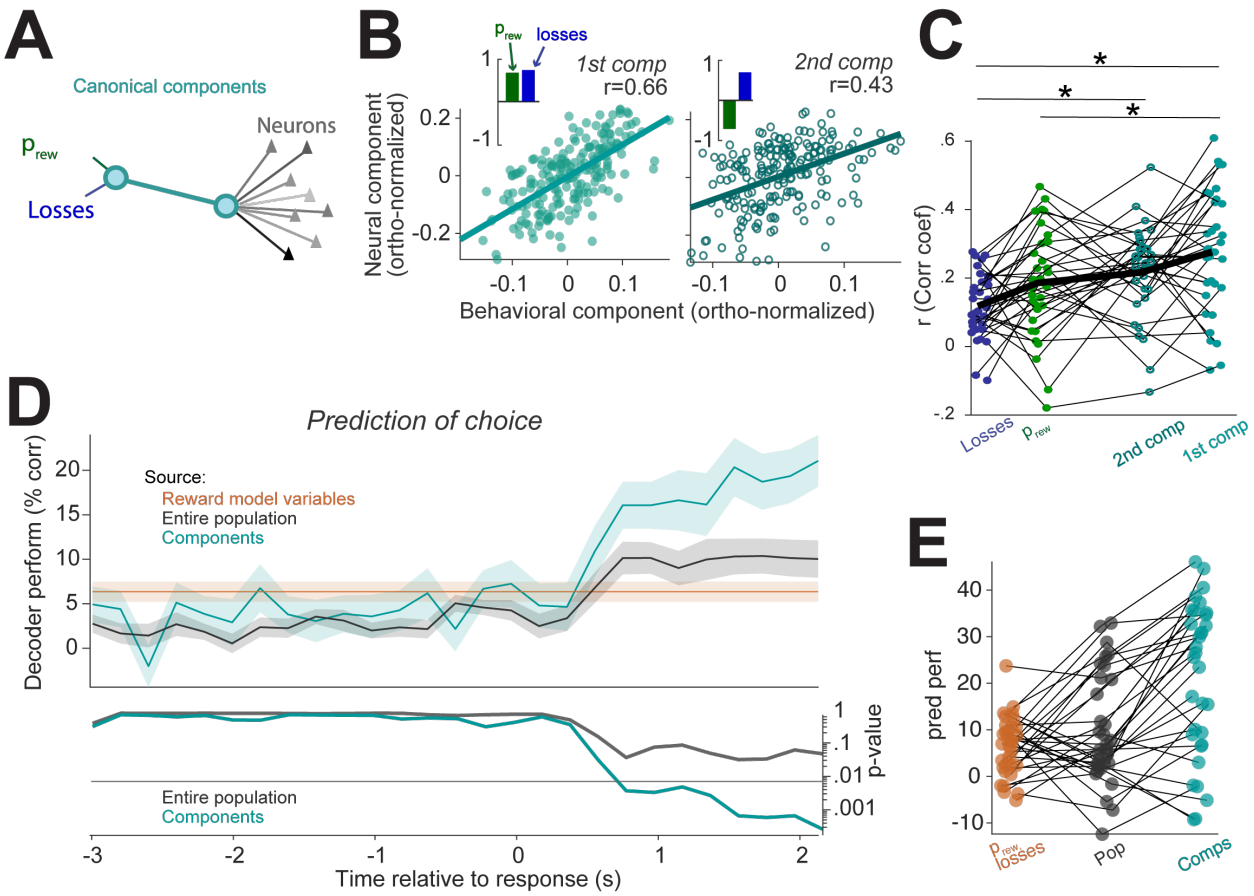

Figure 4 – Canonical components of the neural population representing the variables of the reward model predict decisions better than the entire population. (A) Illustration of the canonical component analysis that relates the variables of the reward model to the activity of a simultaneously recorded population of neurons. Firing rates were averaged within a 1-s interval prior to each behavioral response. **(B)** Scatter of values for the first and second pairs of the canonical components, shown for one session. *Inset*: The coefficients of the two task variables in the associated behavioral component. **(C)** Cross-validated correlation coefficients of the canonical components, compared to the cross-validated correlation coefficient between each variable of the reward model and the individual neuron most correlated with that variable. Each data point represents one session. **(D)** *Top*: Performance of an SVM decoder predicting choice using various sources: the task variables (as in Fig. 2E), the firing rates of the entire simultaneously recorded population within each 200 ms time bin (starting 3 s before and ending 2 s after each behavioral response), or the projections of that population activity onto the canonical components. *Bottom*: Two traces of *p*-values for one-sided Wilcoxon signed-rank test, representing the difference between prediction performance using the neural activity (either the entire population or the canonical components) and prediction performance using the reward model variables. The significance threshold with multiple comparison correction was 0.007, shown as a horizontal gray line. **(E)** Prediction performance for 33 sessions using the last time bin in panel D.

METHODS

All experiments were performed under protocols approved by The University of Texas at Houston Animal Care and Use Committee (AWC) and the Institutional Animal Care and Use Committee (IACUC) for the University of Texas Health Science Center at Houston (UTHealth). Two adult male rhesus monkeys (Macaca mulatta; monkey G: 15 kg, 9 years old; monkey T: 12 kg, 9 years old) were used in the experiments.

Behavioral training and testing

After habituating each monkey with the custom-made experimental cage (4'x2'x3' LxWxH) for at least 4 days per week for over 4 weeks, we trained them to press the knob on each box to receive rewards. Over the course of 4-6 months, we gradually increased the mean time in the VI schedule to let the monkeys grasp the concept of probabilistic reward delivery. Once we started using VI10 (corresponding to an average reward rate of 0.1 rew/s) or higher (less than 0.1 rew/s), monkeys started to spontaneously switch back and forth between the two boxes. If the monkeys disengaged from the task or showed signs of stress, we decreased the VI schedule (corresponding to increasing the reward rate) and kept it constant for one or two days. If the monkey showed a strong bias toward one reward source, we used unbalanced schedules to encourage the monkeys to explore the less preferred box.

After training, we tested monkeys using a range of balanced and unbalanced reward schedules. For balanced schedules we used VI20 or VI30 on both boxes. For unbalanced schedules, we used VI20 versus VI40, VI15 versus VI25, or VI10 versus VI30. The unbalanced schedules may reverse once, twice or three times during a session, e.g. the box with VI20 becomes VI40 and the box with VI40 becomes VI20 after the reversal. Each session lasts until the monkey receives 100 or 200 rewards 1-7 hours including a 1-hour break after 100 rewards in sessions with 200 rewards. If the monkeys were not engaged with the task for more than 2 minutes, we sometimes interrupted them to encourage them to engage with the task. For the analysis, we exclude all responses which occurred more than 60 s or less than 1 s after the previous response. The lower bound on the inter-response interval was imposed to avoid mixing in the event-locked neural activity.

Analysis of the location and the locomotion

To determine the location and locomotion of the monkey, an overhead wide-angle camera was permanently installed in the experimental cage and the video was recorded at an average rate of 6 frames per second. Each frame (Fig. S7, step-1) was post-processed using custom-made Matlab code with six steps. First, the background image was extracted by averaging all frames in the same experimental session, then it was subtracted from each frame (Fig. S7, step-2). The background-subtracted image was then passed through a manually determined threshold to find the dark areas (Fig. S7, step-3). The same image frame was also processed using standard edge detection algorithms (Fig. S7, step-4). The thresholded and edge detected images were then multiplied together, and the result was convolved with a spatial filter, which was a circle

with the estimated angular diameter of the monkey (Fig. S7, step-5). The peak of this filtered image was marked as the location of the monkey (Fig. S7, step-6). Locomotion (velocity) was calculated as the vector difference between monkey locations in consecutive frames divided by their time difference.

Determining the reward availability and calculating p_{rew}

In each time bin of size *dt* = 10 ms, reward became availability at a given box if a sample from a Bernoulli distribution was 1. The probability of this event was *dt*/*VI* where *VI* is the Variable Interval schedule (in seconds). When the reward became available, it stayed available until collected by the animal. This makes the probability of reward availability a function of the scheduled variable interval as well as the time since the last response:

$$p_{rew} = 1 - (1 - dt/VI)^{t/dt}$$

where *t* is the time since the last response in seconds (Fig S1).

Chronic implantation of the Utah array

A titanium head post (Christ Instruments) was implanted, followed by a recovery period (> 6 weeks). After acclimatization with the experimental setup, each animal was surgically implanted with a 96-channel Utah array (Blackrock Microsystems) in the dorsolateral prefrontal cortex (dlPFC) (area 46; anterior of the Arcuate sulcus and dorsal of the Principal sulcus (Figure S8). The stereotaxic location of dlPFC was determined using MRI images and brain atlases prior to the surgical procedure. The array was implanted using the pneumatic inserter (BlackRock microsystems). The pedestal was implanted on the caudal skull using either bone cement or bone screws and dental acrylic. Two reference wires were passed through the craniotomy under and above the dura mater. After the implant, the electrical contacts on the pedestal were protected using a plastic cap all the time except the experiment time. Following array implantation, animals had at least a 2-week recovery period before recording from the array.

Recording and Pre-processing the neural activity

To record the activity of neurons while minimizing the interference with the behavioral task, we used a lightweight, battery-powered device (Cereplex-W, BlackRock Microsystems) that communicates wirelessly with a central amplifier and digital processor (Cerebus Neural signal processor, BlackRock Microsystems). First, the monkey was head-fixed, the protective cap of the array's pedestal was removed, the contacts were cleaned using alcohol and the wireless transmitter was screwed to the pedestal. The neural activity was recorded in the head fixed position for 10 minutes to ensure the quality of the signal before releasing the monkey in the experimental cage which was surrounded by eight antennas. In the recorded signal, spikes were detected online (Cerebus neural signal processor, Blackrock Microsystems) using single (manually selected) or double (±6.25 times the standard deviation of the raw signal) thresholding of the raw electrical activity in each channel. To minimize the recording noise, we optimized the electrical grounding by keeping the connection of the pedestal to the bone clean and tight. The on-site digitalization in the

wireless device also showed lower noise than common wired head-stages. The remaining noise from the movements and muscle activities of the monkeys was removed offline using the automatic algorithms in offline sorting (Plexon inc.). Briefly, this was done by removing the outliers (outlier threshold = 4-5 standard deviations) in a 3-dimensional space that was formed by the first three principal components of the spike waveforms. Then, the principal components were used to sort single units using the expectation-maximization algorithm. Each single and multi-units were evaluated using several criteria: showing consistent spike waveforms, modulation of activity during the 1-sec interval before or after the button pushes, and exponentially decaying ISI histogram with no ISI shorter than the refractory period (1 ms). The analyses used all spiking units with consistent waveform shapes (single units) as well as spiking units with mixed waveform shapes but clear pre- or post-response modulation of firing rates (multi-units).

References:

Aldiss, M., & Davison, M. (1985). Sensitivity of time allocation to concurrent-schedule reinforcement. *Journal of the Experimental Analysis of Behavior*, *44*(1), 79–88. https://doi.org/10.1901/jeab.1985.44-79

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Perspective Not Noisy, Just Wrong: The Role of Suboptimal Inference in Behavioral Variability. *Neuron*, *74*, 30–39. https://doi.org/10.1016/j.neuron.2012.03.016

Bracis, C., Gurarie, E., Moorter, B. Van, & Goodwin, R. A. (2015). Memory Effects on Movement Behavior in Animal Foraging, 1–21. https://doi.org/10.1371/journal.pone.0136057

Bunge, S. A., Burrows, B., & Wagner, A. D. (2004). Prefrontal and hippocampal contributions to visual associative recognition: Interactions between cognitive control and episodic retrieval. *Brain and Cognition*, *56*(2 SPEC. ISS.), 141–152. https://doi.org/10.1016/j.bandc.2003.08.001

Calapai, A., Berger, M., Niessing, M., Heisig, K., Brockhausen, R., Treue, S., & Gail, A. (2017). A cage-based training, cognitive testing and enrichment system optimized for rhesus macaques in neuroscience research. *Behavior Research Methods*, *49*(1), 35–45. https://doi.org/10.3758/s13428-016-0707-3

Chandrasekaran, C., Peixoto, D., Newsome, W. T., & Shenoy, K. V. (2017). Laminar differences in decision-related neural activity in dorsal premotor cortex. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00715-0

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136. https://doi.org/10.1016/0040-5809(76)90040-X

Elliott, R., Friston, K. J., & Dolan, R. J. (2000). *Dissociable Neural Responses in Human Reward Systems*.

Fagot, J., Gullstrand, J., Kemp, C., Defilles, C., & Mekaouche, M. (2014). Effects of freely accessible computerized test systems on the spontaneous behaviors and stress level of Guinea baboons

(Papio papio). *American Journal of Primatology*, *76*(1), 56–64. https://doi.org/10.1002/ajp.22193

Fernandez-Leon, J. A., Parajuli, A., Franklin, R., Sorenson, M., Felleman, D. J., Hansen, B. J., … Dragoi, V. (2015). A wireless transmission neural interface system for unconstrained non-human primates HHS Public Access. *J Neural Eng*, *12*(5), 56005. https://doi.org/10.1088/1741-2560/12/5/056005

Freedman, E. G. (2008). Coordination of the Eyes and Head during Visual Orienting Edward, *190*(4), 369–387. https://doi.org/10.1007/s00221-008-1504-8.Coordination

Gibbon, J., Church, R. M., Fairhurst, S., & Kacelnik, a. (1988). Scalar expectancy theory and choice between delayed rewards. *Psychological Review*, *95*(1), 102–114. https://doi.org/10.1037/0033-295X.95.1.102

Hernandez-peon, R., Scherrer, H., & Jouvet, M. (1956). Modification of electric activity in cochlear nucleus during attention in unanesthetized cats. *Science (New York, N.Y.)*, *123*(3191), 331–332. https://doi.org/10.1126/science.123.3191.331

Herrnstein, R. J., Rachlin, H., & Laibson, D. I. (1997). *The matching law: Papers in psychology and economics by Richard Herrnstein*. Harvard University Press. Retrieved from https://psycnet.apa.org/record/1997-97495-000

Iigaya, K., Ahmadian, Y., Sugrue, L. P., Corrado, G. S., Loewenstein, Y., Newsome, W. T., & Fusi, S. (2019). Deviation from the matching law reflects an optimal strategy involving learning over multiple timescales. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09388-3

Kennerley, S. W., & Wallis, J. D. (2009). Reward-Dependent Modulation of Working Memory in Lateral Prefrontal Cortex. *Journal of Neuroscience*, *29*(10), 3259–3270. https://doi.org/Doi 10.1523/Jneurosci.5353-08.2009

Knight, K. (2011). Head movements give away foraging behaviour. *Journal of Experimental Biology*. Company of Biologists Ltd. https://doi.org/10.1242/jeb.066795

Milton R, Shahidi N, Dragoi V (2019). Dynamic states of population activity in prefrontal cortical networks of freely-moving macaque. *In review*.

Schneider, S. M., & Davison, M. (2005). Demarcated response sequences and generalised matching. *Behavioural Processes*, *70*(1), 51–61. https://doi.org/10.1016/j.beproc.2005.04.005

Smith, J. M. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511806292

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching Behavior and the Representation of Value in the Parietal Cortex, *304*(October), 457–461.

Tollin, D. J., Populin, L. C., Moore, J. M., Ruhland, J. L., Yin, T. C. T., Daniel, J., … Yin, T. C. T. (2019). Sound-Localization Performance in the Cat : The Effect of Restraining the Head, 1223–1234. https://doi.org/10.1152/jn.00747.2004.In

Yin, M., Li, H., Bull, C., Borton, D. A., Aceros, J., Larson, L., & Nurmikko, A. V. (2013). An externally head-mounted wireless neural recording device for laboratory animal research and possible human clinical use. In *Proceedings of the Annual International Conference of the IEEE Engineering in*

Medicine and Biology Society, EMBS (pp. 3109–3114).
https://doi.org/10.1109/EMBC.2013.6610199

Supplementary figures

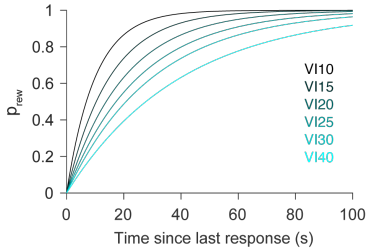

Figure S1 | The probability of reward availability as a function of the scheduled reward rate and the time since last response on the same box.

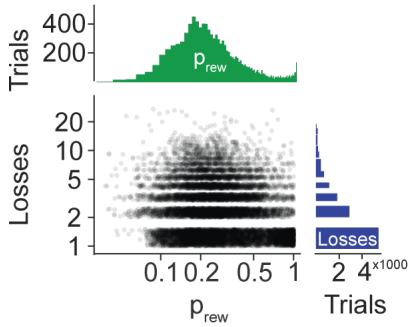

Fig S2 | Joint distribution of the *loss count* and p_{rew} for 13269 responses, and the marginal distributions of p_{rew} and the *loss count*.

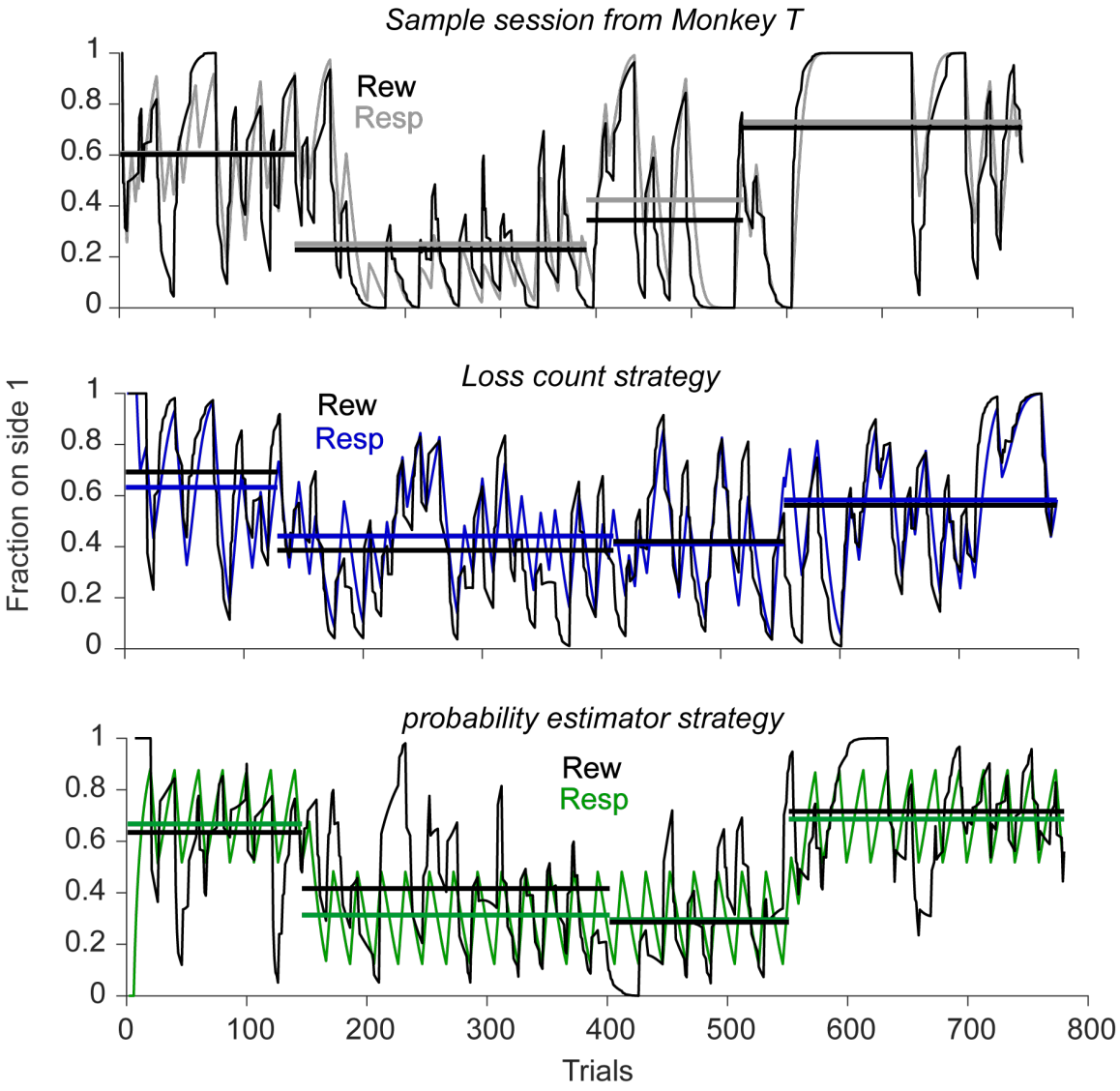

Figure S3 | Dynamic matching for a sample session of Monkey T with 3 sets of reward schedules: VI15-VI25, VI25-VI15, and VI15-VI25 again. We compared two simulated agents, one with loss counting strategy (blue) and the other with probability estimation strategy (green). The reward and response rates were calculated locally using a causal Gaussian filter (Sugrue et al., 2004).

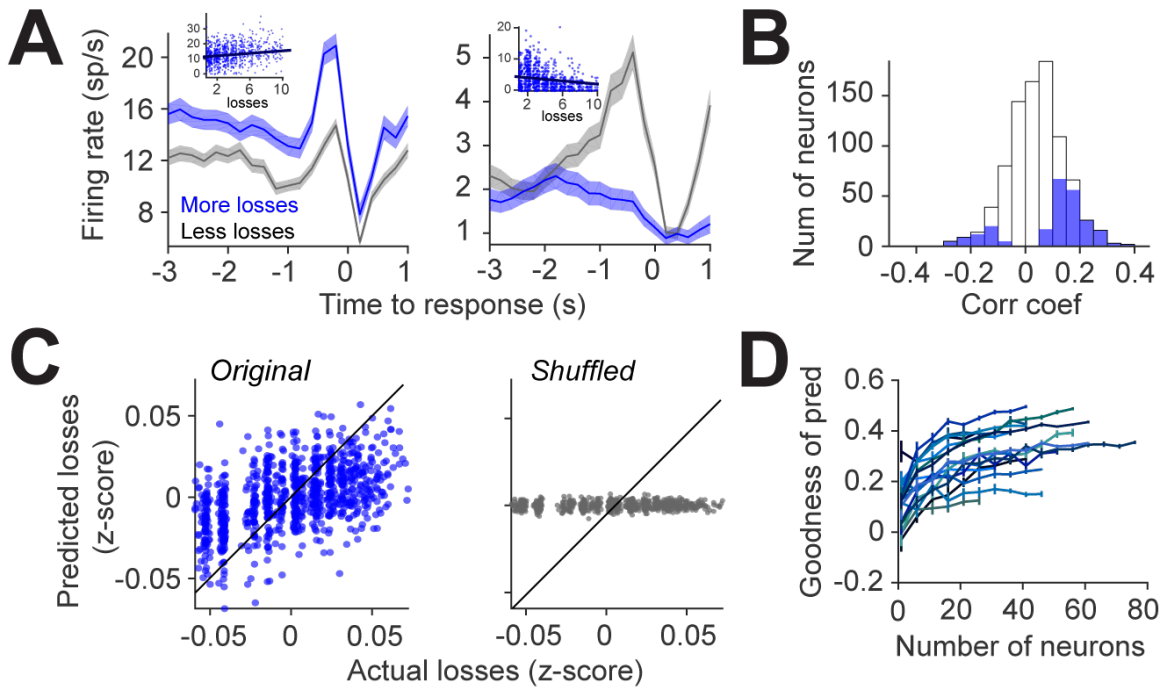

Figure S4 | A similar analysis to figure 3D, G, and H, for *loss count* instead of p_{rew}. (A) Two sample neurons for which the pre-response firing rate modulates the *loss count*. The firing rate was calculated for each 200 ms time bin starting 3 s before and ending 1 s after the responses. Firing rates were averaged over responses with low (< 20th percentile) *loss count* (gray) as well as high (> 80th percentile) *loss count* (blue). *Inset*: For the same neuron, the firing rate of the 1 s interval before each response (-1.1 to -0.1 s) against *loss count*. **(C)** Prediction of the *loss count* in all responses of a sample session, using the pre-response activity of simultaneously recorded neurons in this session. The predictor was a cross-validated regression model with RBF kernel and support vectors. The correlation between the predicted *loss count* on the y-axis and actual *loss count* on the x-axis was used to quantify the goodness of fit. **(D)** Prediction performance for 31 out of 33 sessions as a function of the number of neurons that were used as predictors. The predictor neurons were chosen randomly from the population. The random selection was done 50 times for each data point (starting with 1 neuron at a time, then increasing up to the number of neurons in the session in steps of 5 neurons).

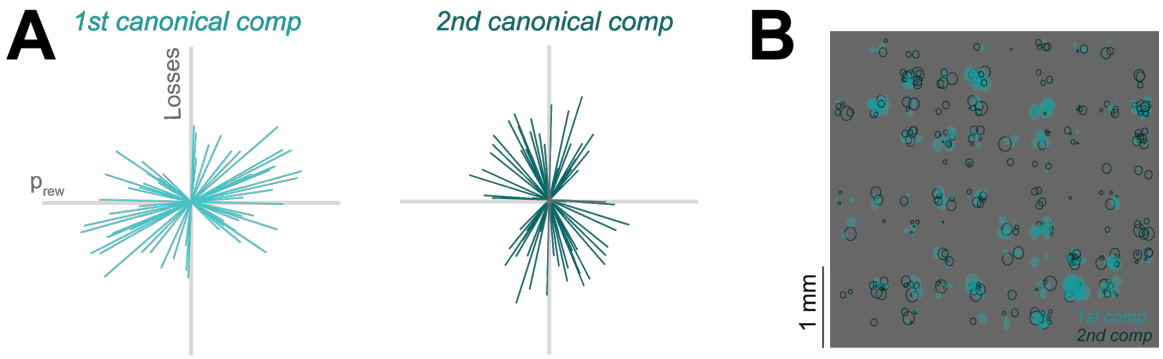

Figure S5 | (A) The placement of canonical components in the 2D space of p_{rew} and *loss count* for 33 sessions. The length of the line segments is proportional to the correlation coefficient between pairs of components in this space and neural space. **(B)** For a sample session, the contribution of each neuron in the first and second canonical components. The diameter of the circles is proportional to the contribution. The location of each circle shows the anatomical location of the recorded neuron on the map of the multi-electrode array.

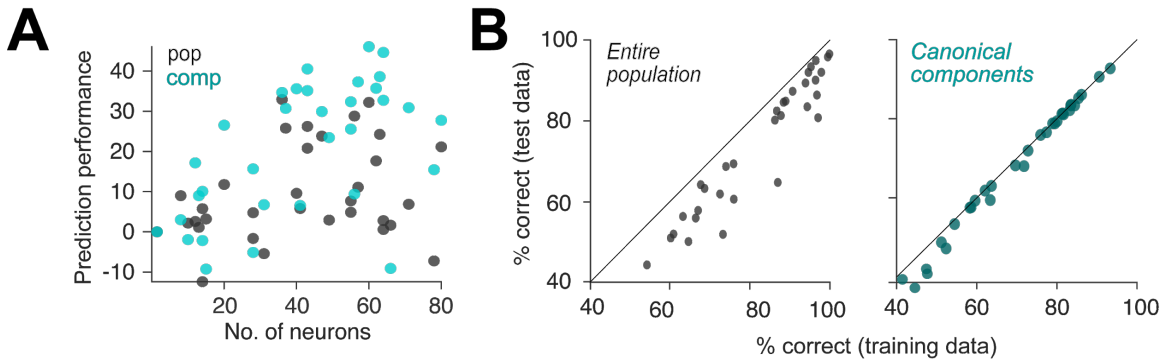

Figure S6 | (A) Prediction performance of the choice using the entire population (gray) or the projection of the population on the canonical components (blue) as a function of the number of recorded neurons in each of the 33 sessions. **(B)** The performance of the choice decoders on the training data vs. the test data. For the decoder that was using the entire population of recorded neurons, the performance on the training data was higher than the test data, suggesting an over-fitting of the decoder parameters. Note that the performances of decoders in this panel are not shuffle-corrected.

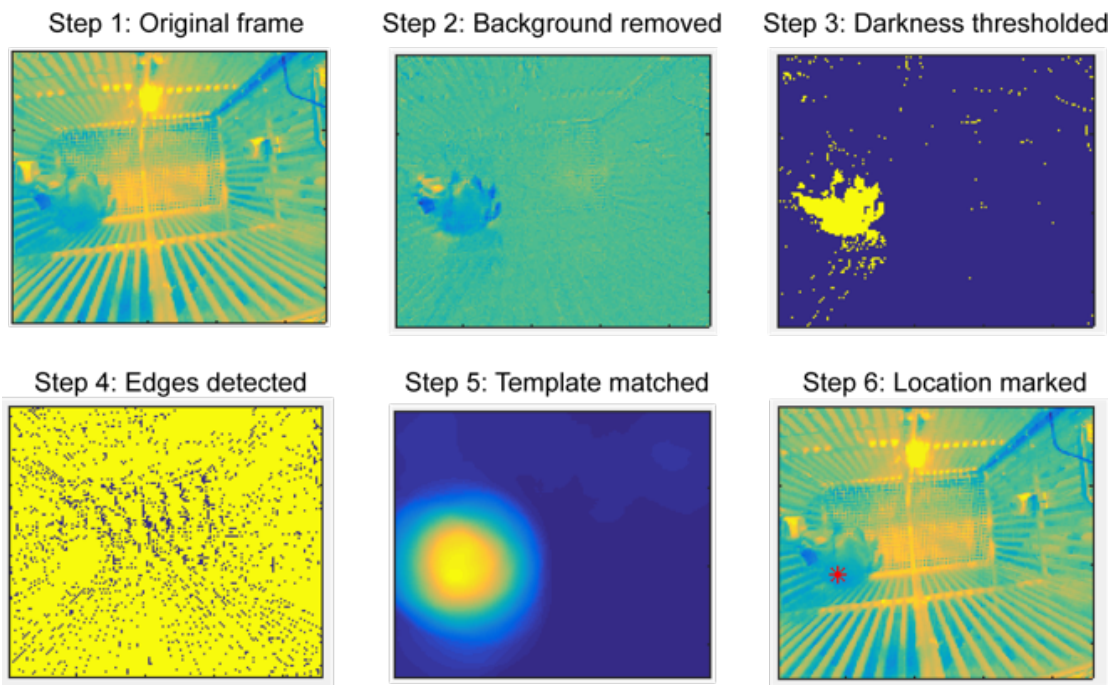

Fig S7 | Processing the images of the overhead camera to locate the monkey. See Methods for the description of steps 1-6**.**

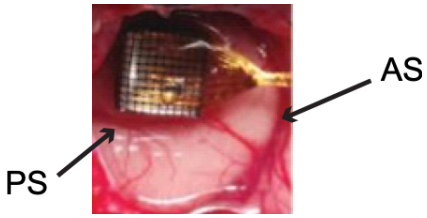

Figure S8 | The location of a 96-channel Utah array in dlPFC on the left hemisphere of monkey G. The Arcuate sulcus (AC) and principal sulcus (PS) are marked.