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

Neda Shahidi[1,2], Paul Schrater[3,4], Anthony Wright[1], Xaq Pitkow[5,6,7*], Valentin Dragoi[1,6*]

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

[2] German Primate Center
Göttingen, Germany, 37077

[3] Dept. of Computer Science
University of Minnesota
Minneapolis, Minnesota

[4] Dept. of Psychology
University of Minnesota
Minneapolis, Minnesota

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

[6] Dept. of Electrical and Computer Engineering
Rice University
Houston, TX 77005

[7] Center for Neuroscience and Artificial Intelligence
Baylor College of Medicine
Houston, TX 77030

* denotes equal contribution and corresponding authors

ABSTRACT: To optimize their foraging strategy, animals must continuously make decisions about where to look for food and when to move between locations of possible food sources. Until now it was difficult to examine the neural bases of foraging in naturalistic environments because previous approaches have relied on restrained animals performing trial-based foraging tasks. Here, we allowed unrestrained monkeys to freely interact with concurrent reward options while we wirelessly recorded population activity in dorsolateral prefrontal cortex (dlPFC). Although the relevant reward dynamics were hidden from the animals, they were nonetheless encoded in the population activity and helped predict foraging choices. Surprisingly, the decoded reward dynamics were represented in a subspace of the high-dimensional population activity, and predicted animal's subsequent choice better than either the true experimental variables or the raw neural responses. Our results indicate that monkeys' foraging strategy is based on a cortical model of reward dynamics as animals freely explore their environment.

While foraging, animals must continuously make decisions about where to search for food and when to move between possible food sources. To survive in environments with sparse resources, animals forage more effectively if they can predict future outcomes before they execute costly actions such as relocation (1–3). Two major limitations of past neuroscience studies of foraging have impeded our understanding of this natural behavior. Specifically, trial-based tasks are unable to expose the continuous decision-making process during food search and selection, and experiments conducted with head restraints may substantially distort behavior and its underlying natural computations.

Existing foraging theories based on decades of behavioral experiments in various species revolve around the idea of melioration, or the matching law (4). This law states that an animal dedicates time or effort to an option in proportion to its value, which is estimated from the recent history of reward delivery. Although tracking reward history does enable the animal to detect changes in reward rate, this average history may not capture the predictable aspects of future rewards, and therefore matching can lead to inefficient foraging (5). Deviations from this matching law may reveal more sophisticated mental computations. Additionally, examining foraging in trial-based tasks makes it difficult to examine the neural bases of the continuous decisions the animal would make freely about when to move between food locations. Therefore, previous foraging studies have mainly focused on predicting the expectation of reward without attempting to explain animal's decisions to stay or switch sides during foraging (6).

Restraining animals to record neural activity can cause other inefficiencies in animal behavior (7,8). The consequences of physical restraints may be especially dramatic on food-seeking behavior because animals typically use environmental information or head movements for foraging (9,10). Furthermore, the dynamics of neural populations have been shown to differ in freely moving animals compared to physically restrained ones (11). Here, we addressed limitations of previous studies by designing a freely-moving foraging task,

thus allowing animals to continuously interact with the task and explore a wide range of reward expectancies. We discovered that unrestrained animals do not simply follow the reward flow, but correctly estimate the instantaneous chance of the next reward and use this estimation to make subsequent choices. We hypothesized that the subjective estimation of reward dynamics can be decoded from the population activity in dorsolateral prefrontal cortex (dlPFC), an area where neural responses encode reward (12–14) and are related to memory (15) and motor planning (16). For decades, recording electrophysiological activity from unrestrained animals has been desired (17) but technologically challenging. We overcame technical challenges by performing wireless recordings from chronically implanted electrode arrays (18,19) while designing an experimental setup for the effective transmission of a low power electromagnetic signal (see Methods) (20,21).

RESULTS

Monkeys (n=2) were exposed to two concurrent reward sources on a variable interval schedule (22). We made it costly for the animal to switch between reward sources by placing them 120 cm apart (Fig. 1A, left). Animals freely interacted with the task equipment, and we did not impose a trial structure or a narrow response window (see Methods). A multi-electrode Utah array was chronically implanted in dorsolateral prefrontal cortex (dlPFC) while spiking activity was collected using a light-weight, energy-efficient wireless device (Fig 1A, right and Fig. 1B) (19).

Rewards on both sides (box 1 and box 2) became available at exponentially distributed random times after a previous reward, with possibly different mean times or 'schedules' (*i.e.,* constant hazard rates; Fig. 1C). Schedules were chosen to be 10, 15, 25, or 30 seconds and were constant for a block of rewards. Multiple schedules allowed us to diversify the response dynamics (22). Each experimental session contained 2–4 blocks with 34 or 66 rewards in each block. Once becoming available, each reward remained available until the animal pressed a knob (a "response"), at which time it was delivered (Fig. 1C). Prior to delivery, reward availability was hidden from the monkey. Given the constant hazard rate and no reward disappearance, the *probability* of reward availability, p_{rew}, increased exponentially toward 1 with the time elapsed since the last response, with a time constant that depended on reward schedule (see Methods; Fig S1). As the monkey chose when to respond, its decisions influenced p_{rew} (Fig. 1C). Although p_{rew} completely determines the chance of receiving a reward for the subsequent response within a block (Fig 1D), animals may choose to weigh the recent history of reward (melioration). For example, tracking the *loss count* (Fig. 1C) — the number of consecutive unrewarded responses — does not determine the chance of receiving a reward if the schedule is known, although it could in principle be useful while learning an unknown schedule. However, in practice we found that the loss count does not appreciably predict rewards experienced by the animal (Fig. 1E). According to the marginal value theorem of foraging theory (1), an animal could optimize its reward while minimizing travel costs by tracking the temporal evolution of p_{rew} and using it to choose

when to search for reward and when to switch sides. We hypothesized that foraging animals estimate p_{rew} and track the temporal evolution of *loss count* to extract the underlying reward dynamics.

The neurons' response-locked events, *i.e.*, firing rates 1 s before and after each response (Fig. 1B) were extracted from the spiking activity of the multiple cells recorded simultaneously (monkey G: 14 sessions; monkey T: 19 sessions, n = 1,405 single and multi-units). We reasoned that this interval is sufficiently long to capture the relevant neural events that encode task-relevant variables, and relate spike rates to behavioral and task events. The average firing rates of a sample neuron (Fig. 1F) is shown for three consecutive box responses with various reward/choice outcomes: unrewarded responses when the animal stayed on the same box, unrewarded responses followed by a switch to the other box, and rewarded responses when the animal stayed on the same box. The fourth condition, switching to the other box after a rewarded response, accounted for only 2% of the responses, hence we ignored it. Before attempting to understand which neural events determined the choice to stay or switch locations after unrewarded responses, we first examined whether reward history, quantified as *loss count* and the true probability of reward availability, p_{rew}, predict animal's choice.

Predicting choice using reward dynamics

Animals of various species often display matching behavior, i.e., choosing between concurrent reward sources proportional to their value (4). However, matching, while ubiquitous, does not entail a single computational strategy. For our foraging task, matching behavior is consistent with substantially different strategies. One strategy can be simulated using an agent that switches to the other box after the *loss count* exceeds a noisy threshold (Fig 2A). This strategy corresponds to a basic 'win-stay / lose-switch' rule. We implemented this strategy by sampling the threshold from a Gaussian distribution with the same mean and variance as the *loss count* distribution at times when the animal switches sides. Although this agent is blind to both the average reward rate and the probability of the next reward, it still follows the generalized matching law (Fig. 2B and Fig. S3). The slight under-matching that we observed resembles the behavior of various species in previous studies (4).

We examined a more complex strategy that tracks reward probability and uses foraging theory to make choices by involving three variables: (i) the time since the preceding response and (ii) the variable-interval schedule — which together determine the probability of reward — and (iii) the relative cost of switching locations, which affects the threshold for when to switch. We simulated an agent that follows such a strategy by making choices based on the correct 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. 2A), in accordance with the Marginal Value Theorem of foraging theory (1). 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. 2B and Fig. S3). These two simulations show that

an approximate matching law may arise when following a strategy that is either blind to timing, or fully informed. This implies that matching behavior is not, by itself, informative about the underlying strategy or animals' ability to grasp the hidden rule of the task.

We subsequently examined how individual choices depend on context to determine whether the animal's strategy is sensitive to the hidden rules of the task. After each attempt to gather reward by pressing the knob, we predicted the animal's choice to stay on the same box or switch sides. We based the prediction on three variables of the reward dynamics: the recent history of reward delivery (*loss count),* the probability of reward (p_{rew}), and the reward outcome. In 98% of rewarded responses, monkeys chose to stay at the same box, and hence we did not analyze rewarded responses further. However, for unrewarded responses, the chance of switching varied with both p_{rew} and *loss count* (Fig. 2C). Trained nonlinear binary classifiers (support vector machines with radial basis functions) predicted behavioral decisions better than chance (Fig. 2D), using p_{rew} (6.3% better than chance, $p = 0.006$, Wilcoxon signed rank with multiple comparison correction) but not from *loss count* (0.9% above chance level, $p = 0.48$). This is surprising since p_{rew} is completely hidden from the animals while *loss count* is observable. Predictions were best when using both variables (7.7% above chance, $p < 0.001$; 8% above chance for monkey G, $p = 0.01$; 5% for monkey T, $p = 0.01$). Thus, when a response is unrewarded, the recent history of past rewards and probability of the reward availability together help predict the animal's next choice.

Decoding reward dynamics from the neural population

Given that our two key task variables, p_{rew} and *loss count*, provide information about the animal's upcoming choices, we further examined whether they are encoded by the dlPFC neural population. The responses of neurons were only weakly correlated with animal's movement-related variables (Fig. 3A–B), such as spatial location within the environment and locomotion (location *X*: $r = 0.17$, Pearson correlation, $p \ll 10^{-3}$, location *Y*: r = 0.12, $p = 0.03$, and locomotion: $r = 0.08$, $p \ll 10^{-3}$). Furthermore, our control experiments revealed that eye movements have only a minor influence on neuronal responses while the animal interacts with the box, although they have more influence during locomotion ($r = 0.16$, $p \ll 10^{-3}$, for eye velocity, and $r = 0.13$, $p \ll 10^{-3}$, for fixation rate (20)). We thus focused on reward-related variables by removing from neural activity the components corresponding to spatial location (*loc X* and *loc Y*) and locomotion (*loc D*, see Methods) by subtracting its vector projection onto a three-dimensional subspace spanned by three task-irrelevant variables (Fig. 3B-C; *loc X*, *loc Y* and *loc D* (Methods). The remaining neural activity was uncorrelated with these task-irrelevant variables (Fig. 3A-C).

For each neuron ($n = 1405$ single and multi-units), we measured the spike counts in a 1-second interval before each behavioral response (*i.e.*, the "pre-response" interval from –1.1 to –0.1 seconds). This time interval was selected since the arm movement starts approximately 0.5 s before the response is recorded, and the modulation of neural activity typically starts around 0.5 s before that movement (23). Consistent

with previous reports (6,12,14,24), we found that dlPFC neurons encode aspects of the recent history of reward (or, more precisely, the lack thereof) through *loss count*. Figs. 3D and E show an example neuron whose pre-response firing rate was significantly higher after reward (*loss count* = 0) than after a few unrewarded responses (*loss count* ≥ 2, Wilcoxon rank-sum test, $p \ll 10^{-3}$). Over the entire population, there was a significant correlation between the pre-response firing rate and log(*loss count*+1) (*t*-test, $p < 0.01$) for 19.4% of the neurons (Fig. 3F: 15.2% positively correlated and 4.2% negatively correlated; Monkey G: 9.3% of neurons, Monkey T: 23.1% of neurons).

To further examine how information about *loss count* is distributed across cells, we decoded population activity prior to each behavioral response using the spike counts of randomly subsampled sets of neurons (Fig. 3G and Fig. S4; see Methods describing the regression-based decoder analysis). Decoder performance was higher than chance for populations of more than 6 neurons ($p < 0.001$), but not for individual cells ($p = 0.99$). This is consistent with previous studies reporting that the history of reward delivery is represented in prefrontal cortex (14). However, as shown in Fig. 1E, *loss count* does not predict reward outcome. Therefore, we tested if the pre-response neural activity is informative about p_{rew}, since such cells should predict the upcoming reward better than neurons whose pre-response activity encodes *loss count* but not p_{rew}. Fig. 3H shows an example neuron whose mean activity just before a knob press was significantly higher for the top 20% of p_{rew} values than for the bottom 20% of values (Wilcoxon rank-sum test, $p \ll 10^{-3}$). Across trials, the neuron's pre-response firing rate was correlated with log(p_{rew}) (Pearson correlation coefficient = 0.24). For the entire population, around 25% of neurons exhibited a significant Pearson correlation (Fig 3J, *t*-test, $p < 0.01$; 21.1% positively correlated and 3.6% negatively correlated; Monkey G: 15.1% of neurons; Monkey T: 28.2% of neurons). Our decoder analysis revealed that even random neural subpopulations encode p_{rew} (Fig. 3K; $p = 0.78$ for individual neurons; $p \ll 10^{-3}$ for random subsets of > 6 neurons). Overall, neural populations carry more information about p_{rew} than about *loss count* (Wilcoxon rank sum test with multiple comparison correction; $p \leq 0.001$ for 32 out of 33 sessions).

Given that p_{rew} and *loss count* are only weakly correlated (Fig. S2), decoding these variables individually will not clarify whether they are both represented in the neural population activity or if the representation of *loss count* is merely an artifact of its correlation with p_{rew}. We addressed this issue by using Canonical Correlation Analysis (CCA, see Methods) to find components of the neural population activity that are most correlated with linear combinations of the reward dynamics (Fig. 4A). Two pairs of canonical components contained contributions from p_{rew} and *loss count* (Fig. 4B and Fig. S5A), showing that they are accurately encoded in the population activity. Reward dynamics were more correlated with the canonical components of population activity (cross-validated and averaged across sessions, $r = 0.28$ for the first component and $r = 0.21$ for the second component) than with the responses of individual neurons ($r = 0.19$ for p_{rew}, and $r = 0.12$ for *loss count*, Fig. 4C), indicating that both task variables are encoded in neural activity. Moreover, Fig S5B shows the weight with which each component is present in each neuron; these weights are widely

distributed across neurons, indicating that the task information is distributed across the population of cells rather than being segregated into subpopulations.

Decoding choice using the neural representation of reward dynamics

Next, we investigated whether the representation of reward dynamics in dlPFC predicts an animal's actions, particularly its choice to stay or switch after an unrewarded behavioral response. Since the animal cannot know the true latent reward dynamics, its choices can only be driven by its subjective beliefs about their values rather than the objective truth from the experiment. For example, if at some time the monkey overestimates p_{rew} before generating a response at one box that goes unrewarded, he could be more likely to switch to the other box than if he had known the true (lower) value of p_{rew}, or if he had underestimated p_{rew}. Therefore, we predicted the animal's choice to switch sides using the previously calculated canonical representation of task variables, which we interpret as estimates of animal's subjective beliefs. Specifically, we decoded the neural activity by projecting the population response in a given time bin onto the subspace formed by these canonical neural components. These projections were subsequently used as coefficients of the canonical task components to produce subjective estimates of p_{rew} and *loss count* (Fig. A-C). These estimates were used to predict whether the animal would stay or switch after an unrewarded response (Fig. D-F). We found that the task variables derived from the brain's canonical components are better predictors of choice than the corresponding ground truth task variables (Fig. 4E). Indeed, prediction performance was 21% above chance for canonical components compared to 7.7% above chance for the true task variables (Fig. 4F; Monkey G: 7.6%, Monkey T: 32.2%; when compared to the prediction performance of the true task variables: p=0.01 for monkey G, p=10^{-3} for monkey T and $p \ll 10^{-3}$ for all sessions combined).

It might seem obvious that neural features should be better predictors of choice than experimental variables: after all, the animal's brain is making its choice and not the experimental equipment. However, it is not evident *a priori* whether the relevant neural representations would be found within our recorded dlPFC population, nor whether we record enough neurons to capture enough of the animal's choice-relevant information. And even if dlPFC does contain the choice-relevant signals, it is not obvious that the neural components for our specific hypothesized task variables would be the right ones to predict the choices.

It is thus fascinating that these neurally decoded reward dynamics predict stay/switch choices (21% correct guesses) significantly better than the task variables from which they are derived (7.7%, Wilcoxon signed rank test, p<0.001), and better than the full neural population (10%, Wilcoxon signed rank test, p<0.001) (Fig. 4D). Evidently, our analysis identifies a neural subspace containing correlates of latent variables that are relevant for subsequent choices. This subspace also tends to avoid neural dimensions that contain choice-irrelevant variability, since if present these variations would contribute to overfitting and would only hinder our ability to predict choice. We conclude that we are capturing the neural correlates of the animals' subjective beliefs about latent reward dynamics that inform their choices.

DISCUSSION

We used a trial-free, unrestrained-animal approach to demonstrate that freely moving monkeys base their foraging strategy on an internal prediction of reward. This prediction is not based solely on the recent history of reward, but relies on an internal estimation of the probability of reward availability. Indeed, we found that neural populations in prefrontal cortex contain information about recent reward history and probability of reward availability. Complementary to previous research in restrained animals (6,12), we revealed that neural signals not only encode reward information, but also significantly predict animal's choices after each behavioral response during foraging. These findings challenge and extend long-standing theories of reward-seeking behavior (4) that suggest that animals maximize the recent rate of reward, based on the matching law, without constructing a reward model to predict future behavior.

Surprisingly, we found that the targeted 2-dimensional representation within the high-dimensional space of neural population activity predicts choice better than either the behavioral dynamics or the entire population of recorded neural responses. This is an important confirmation of how targeted dimensionality reduction can reveal neural computations better than behavior or unprocessed neural activity. This type of analysis is essential in natural experiments where task variables are correlated.

One limitation of our findings is the extent to which our results can be generalized across other types of reward dynamics. The reward dynamics in our task are stochastic and time-based, and they resemble the repletion of food resources found in nature. Follow-up studies are needed to determine whether our findings apply to other reward schemes, such as non-Markovian, more clock-like dynamics, or those based on response rate (25) whereby reward becomes available after a variable number of responses rather than a variable time-interval.

Additionally, our study does not determine if the enhanced modulation of neural activity prior to a response is due to a representation of a higher reward expectation or a vigorous motor action due to a higher reward expectation. Comparing the neural modulation between simultaneously recorded populations in dlPFC and pre-motor cortex while continuously measuring the vigor of each response can dissect these two representations.

Finally, by allowing animals to move freely during foraging, our study represents a pioneering move toward studying neural correlates of natural cognition in a free-roaming setting. This paradigm shift has been suggested decades ago (17), but is only feasible now due to advances in low-power, high-throughput electrophysiological devices as well as large-scale computing (26). Previous studies have underestimated the cognitive capacity of monkeys during foraging, and this might have primarily been because of their restrictive experimental paradigms. Our freely-moving experimental paradigm likely increases the engagement of natural decision-making processes in the animal's brain, and reduces the distortions in population dynamics associated with unnatural head-fixed tasks (27). The free-roaming setting also enabled us to implement a natural switching cost between two reward options by 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 (20) enhances their cognitive ability to perform the task. Overall, we argue that a shift toward more natural behavior is inevitable for understanding neural mechanisms of cognition (26,28–32).

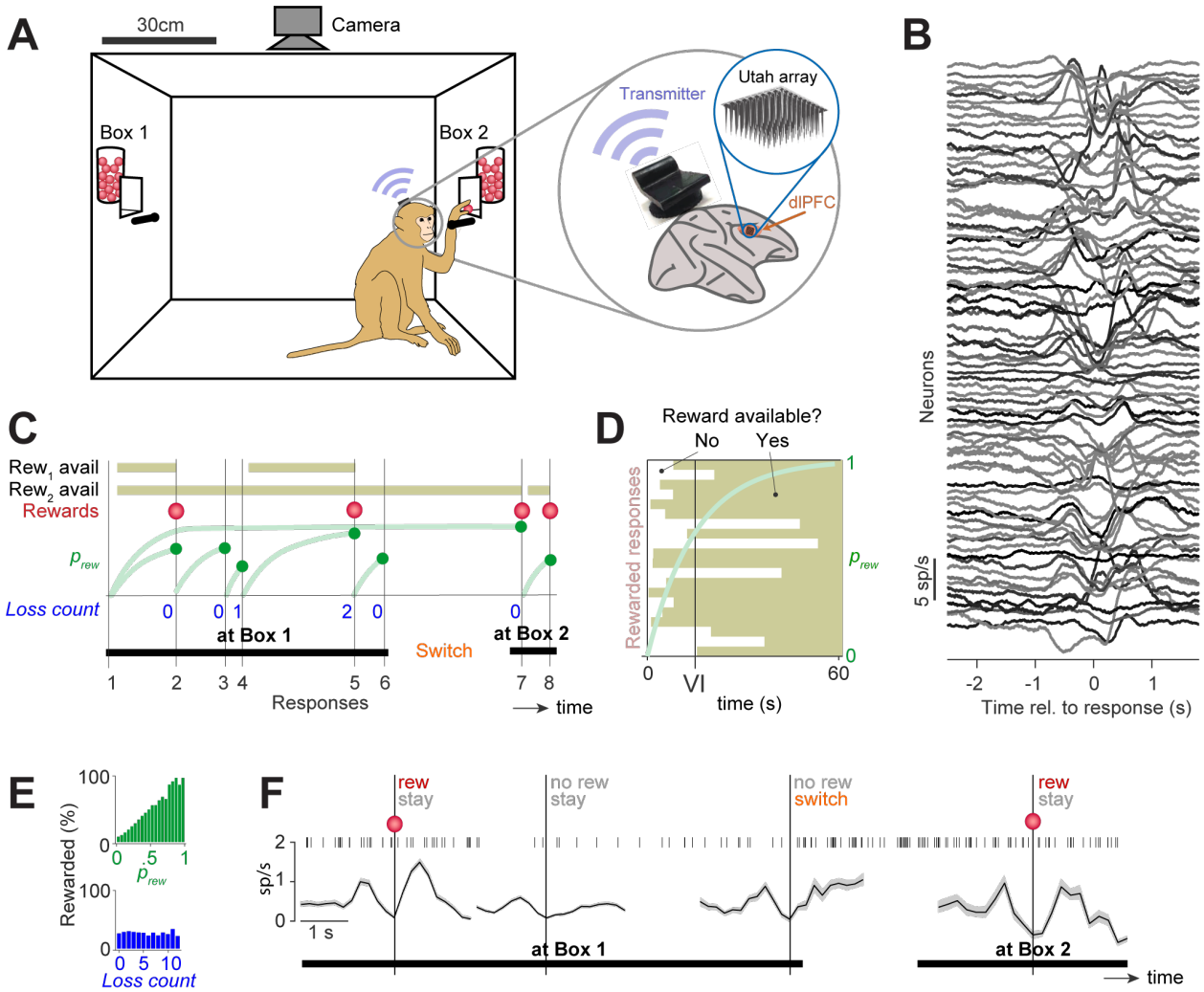

Figure 1 – Foraging in freely-moving monkeys while population activity in prefrontal cortex is recorded wirelessly. (A) *Left*: Schematic of the experimental setup with two reward boxes, two response knobs, and an overhead camera. *Right*: Location of the Utah array in dlPFC (area 46) and wireless transmitter. **(B)** Response-averaged firing rates of 80 single and multi-units recorded simultaneously. **(C)** Illustration of task dynamics with 8 hypothetical behavioral responses (vertical lines) 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 telegraph processes determining the reward availability at boxes 1 and 2. In the example shown, response numbers 2, 5, 7, and 8 were rewarded (third row, red). The time dependence of probability of reward availability is shown in the fourth row (see panel D for a different representation). The value of p_{rew} at the time of each response (green circles) was used in the analyses. The *loss count* (fifth row, blue) is defined as the number of consecutive unrewarded responses at the current box before the current response. **(D)** The monotonically increasing relationship between the time of the rewards and their preceding responses (shown as yellow-green bars in panel C) and the probability of reward availability (see methods; Fig. S1), shown as a horizontal histogram. **(E)** Histogram of 13269 rewarded and unrewarded responses from 33 sessions, showing the percentage of rewarded as a function

of the *loss count* and p_{rew}. **(F)** The spike train of one example neuron on the timescale of four consecutive knob responses showing a variety of events (top row). Event-locked average firing rates of the same neuron are shown in the bottom row, for conditions with reward/no reward and stay/switch choices. For ease of visualization this example used a neuron with a relatively low firing rate compared to others in the population (compare to B).

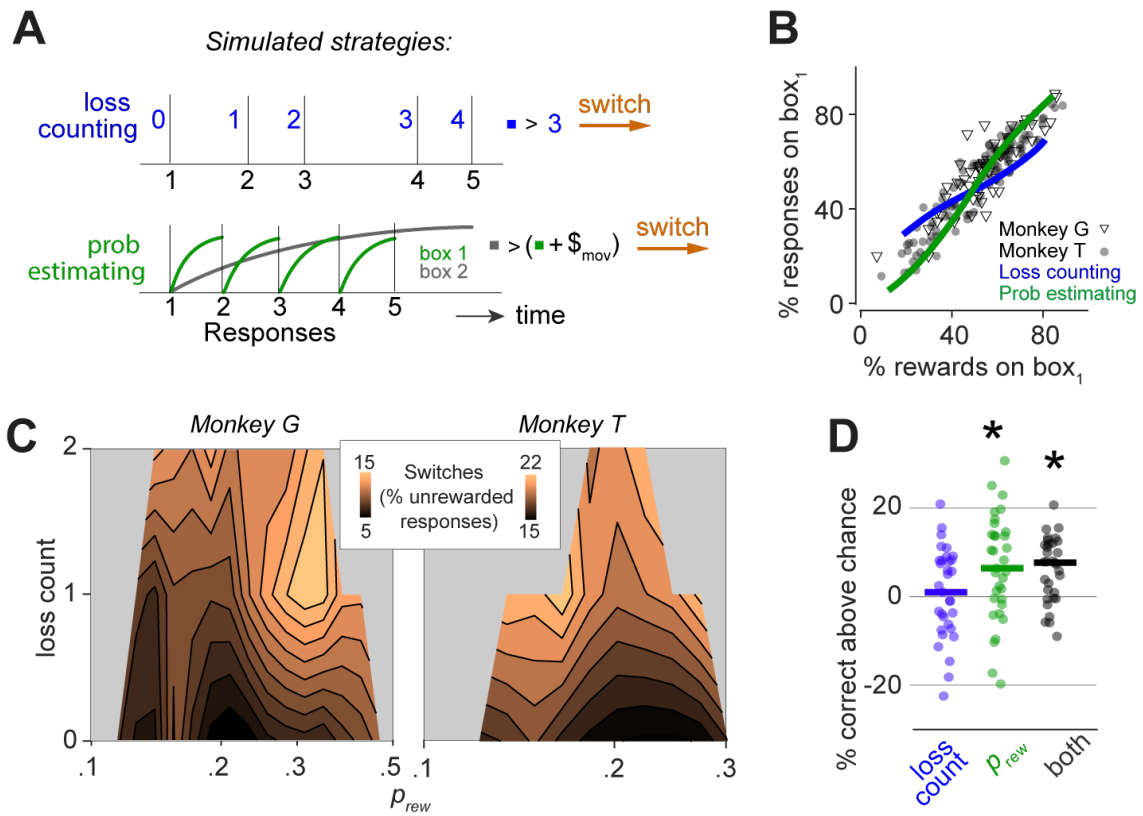

Figure 2 – Behavioral and neuronal performance during foraging. (A) Illustration of foraging strategies for two simulated agents. The '*loss counting*' agent switches to the other box when *loss count* exceeds a threshold drawn from a Gaussian distribution (λ=2.66, σ=1.9). The 'probability estimator' agent switches to the other box when the probability of reward availability on the other box exceeds p_{rew} by a fixed switching cost (1). 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 for each set of reward schedules for box_1 and box_2. 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. **(B)** Matching behavior of two monkeys and two simulated agents: the fraction of behavioral responses at box 1 is approximately proportional to the fraction of rewards at box 1 (26 sets of schedules for monkey G and 59 sets of schedules for monkey T are shown). Curves show spline fits for the matching behavior for the simulated agents. **(C)** 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 of each monkey, sorted according to p_{rew} and smoothed with a window of 300 responses. Responses with *loss count* > 2 were ignored for this analysis since they had < 300 responses. **(D)** The prediction of choice at the time of the response, as the excess percentage of correctly predicted choices beyond chance, using a support vector machine with radial basis function kernels, trained and cross-validated for each session (n=33). Predictors were log(*loss count*+1), log(p_{rew}), or both. We analyzed only unrewarded responses because monkeys switch less than 2% of the time after a reward. 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.

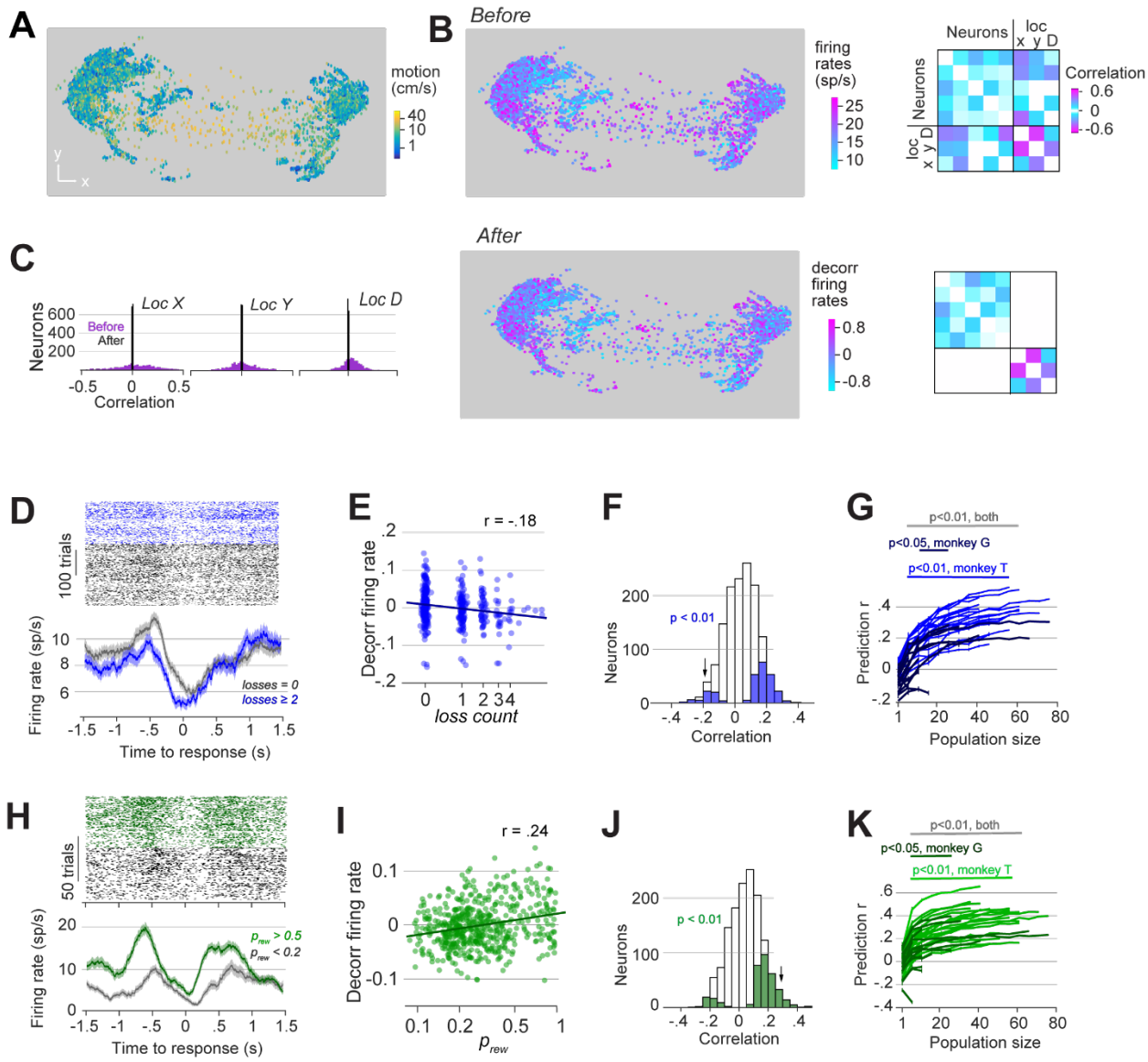

Figure 3 – Neuronal populations encode p_{rew} and *loss count*. (A) Monkey locomotion in a sample session. Only the response-locked time bins (< 3 s before, or < 1 s after a response) are shown. Each dot represents the animal's position in space, sampled every 200 ms. **(B)** *Left*: Population-averaged firing rates for each time bin in (A) before and after subtracting the vector projection of location and locomotion. *Right:* Correlation coefficients between task-irrelevant movement variables (locations Loc X and Loc Y, and locomotion Loc D, of the monkey) and the pre-response firing rate of each neuron, before and after subtracting the vector projection of location and locomotion. For a better 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 covaries with *loss count*. The firing rate was calculated for a 200 ms sliding windows starting 1.5 s before and ending 1.5 s after responses. Firing rates were averaged over responses with low *loss count* (< 20th percentile, gray) and high *loss count* (> 80th percentile, blue). **(E)** for the neuron in D, the scatter of firing rates as a function of the loss count on a log scale for all trials. Because the firing rates were decorrelated from the task-irrelevant variables using a procedure that was explained

in the main text, the values are unitless. **(F)** The distribution of Pearson correlation coefficients between the pre-response firing rate of each neuron and log(*loss count*+1). All neurons from 33 sessions were pooled. The arrow shows the bin containing the example neuron in D and E. **(G)** Prediction of *loss count* for 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 20 times for each data point. 14 sessions of monkey G are shown with a darker color and 19 sessions of monkey T are shown with a lighter color. The matching dark and light blue bars at the top show the range where decoders have performance that differs significantly from chance for the sessions of each monkey. The gray bar shows the significance across all sessions. **(H-K)** Same as D-G but for p_{rew} instead of *loss count*.

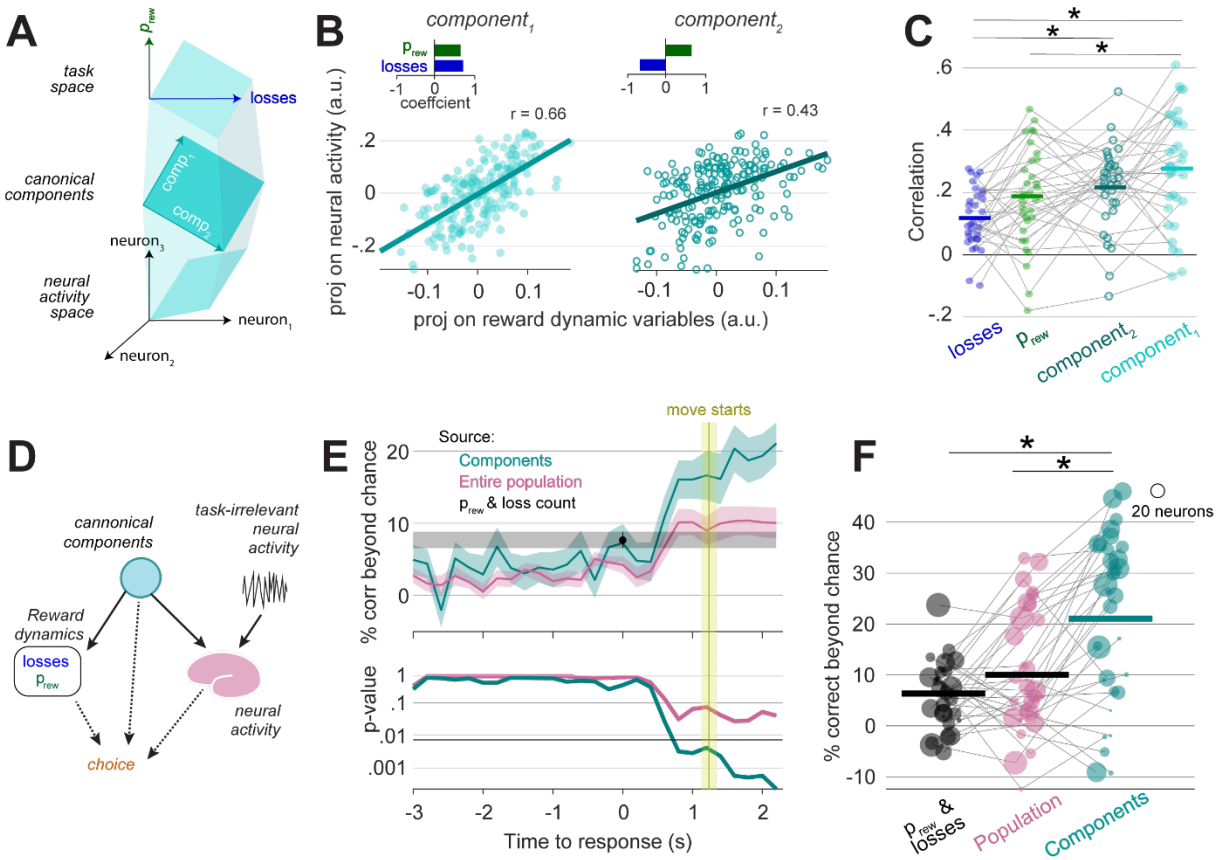

Figure 4 – Canonical components of the neural population representing the variables of the reward model predict decisions better than the entire population. (A) Illustration of the canonical correlation analysis for finding a reduced-dimensional space in the task space and the corresponding subspace in the neural activity space. Two canonical components, $comp_1$ and $comp_2$, define the maximally correlated subspaces between the task variables and the neural activity. **(B)** Scatter plot 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. Stars represent *p*-values < 0.001 after multiple comparison correction. **(D)** Illustration of the canonical component analysis that relates the variables of the reward model to the activity of a simultaneously recorded population of neurons. The cause of the choices is the neural activity which is composed of task-relevant and task-irrelevant components. The task-relevant components are maximally correlated with combinations of the true task variables. **(E) Top:** Performance of an SVM decoder, as the excess percentage of correctly predicted choices beyond chance, 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 (for 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. **(F)** Prediction performance for 33 sessions using the last time bin in panel D panel, for the population and the components. The diameter of the dots represents the number of neurons in each session.

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 in a custom-made experimental cage (120 cm x 60 cm x 90 cm, L×W×H) for at least 4 days per week for over 4 weeks, we trained animals to press the knob on each box to receive a reward. 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), 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 (increased 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, ranging from 1-7 hours including a 1-hour break after 100 rewards in sessions with 200 rewards. If 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 2 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 location dependency and locomotion

To determine the physical 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. S6, step-1) was post-processed in six steps using custom-made Matlab code. First, the background image was extracted by averaging all frames in the same experimental session, then it was subtracted from each frame (Fig. S6, step-2). The background-subtracted image was then passed through a manually determined threshold to identify the dark areas (Fig. S6, step-3). The same image frame was also processed using standard edge detection algorithms (Fig. S6, 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. S6, step-5). The peak of this filtered image was marked as the location of the monkey (Fig. S6, step-6). We used this heuristic because the illumination of the experimental room and the configuration of object was constant. We expect novel techniques for motion and posture detection using deep neural network (33,34) to yield similar results. 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 available 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. 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 preceding response:

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

where *t* is the time since the preceding response (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 S7). 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 at all times except during the experiment. Following array implantation, animals had at least a 2-week recovery period before we recorded from the array.

Recording and pre-processing of 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. The cage was surrounded by eight antennas. In the recorded signal, spikes were detected online (Cerebus neural signal processor, Blackrock Microsystems) using a manually

selected upper threshold on the amplitude of the recorded signal in each channel or an upper and a lower threshold which were ±6.25 times the standard deviation of the raw signal. 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 digitization 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-unit signal was evaluated using several criteria: consistent spike waveforms, modulation of activity with 1-sec of the knob 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).

Removing task-irrelevant components from neural activity

For each neuron k we remove movement-related temporal components of the response r_{kt}, by subtracting its projection onto the subspace spanned by the task-irrelevant variables: $r_k^{\perp} = r_k - \Pi r_k$, where Π is the projection matrix $\Pi = L(LL^{\top})^{-1}L^{\top}$ and L is the T×3 matrix describing the time series of locX, locY, and locD.

Regression-based and binary decoder analysis

To decode a binary variable, such as the choice to stay or switch, we used a support vector machine as a binary classifier with a radial basis function (RBF) kernel. We compared three types of predictive features: task variables (Fig 2E), neural population activity (Fig 4D and E), or canonical components (Fig 4D and E). To decode continuous-values variable such as p_{rew} (Fig 3H) or an integer variable such as the *loss count* (Fig 3I), we used a linear regression model (35). To train these decoders we used 80% of the responses in a given session. The performance of the decoder was measured as the percentage of correctly classified responses for the binary classifier and the Pearson correlation coefficient between the predicted and true values for the regression model. The cross-validated performance for either of the decoders was calculated using the remaining 20% of the responses. This procedure (decoder training and testing) was repeated 100 times, each time using a random subset of the corresponding responses for training and the rest of them for testing. As a control, we trained decoders with randomly shuffled class labels for the binary classifier, or randomly shuffled variables for the regression model (100 random shuffles). The performance of the shuffled decoders was subtracted from the performance of the original decoders and was also used as a null hypothesis for the statistical test of decoder performance.

Canonical correlation analysis (CCA)

Canonical components were calculated using singular value decomposition of the covariance of matrix X in which the columns were log(p_{rew}) and log(*loss count*) and matrix Y in which the columns were the pre-response firing rates of simultaneously recorded neurons, for responses within a session. We used custom code in Matlab to call the functions in R (36) to calculate canonical components and their correlation coefficients. The cross-validation procedure was the same as for the SVM decoder and the regression model.

Statistical analysis

We used the two-sided Wilcoxon signed-rank test except where indicated. We chose this test rather than parametric tests, such as the *t*-test, for its greater statistical power (lower type I and type II errors) when data are not normally distributed. When multiple groups of data were tested, we used the FDR multiple comparison correction for which the implementation was available as a standard function in Matlab. No statistical methods were used to predetermine sample sizes. However, the size of our dataset and the number of the experimental sessions greatly exceeded similar studies.

References:

1. Charnov EL. Optimal foraging, the marginal value theorem. Theor Popul Biol. 1976;9(2):129–36.
2. Smith JM. Evolution and the theory of games. Am Sci. 1976;64(1):41–5.
3. Kilpatrick ZP, Davidson JD, Hady A El. Normative theory of patch foraging decisions. bioRxiv. 2020;
4. Herrnstein RJ, Rachlin H, Laibson DI. The matching law: Papers in psychology and economics by Richard Herrnstein. Harvard University Press; 1997.
5. Sakai Y, Fukai T. The actor-critic learning is behind the matching law: Matching versus optimal behaviors. Neural Comput. 2008;20(1):227–51.
6. Sugrue LP, Corrado GS, Newsome WT. Matching Behavior and the Representation of Value in the Parietal Cortex. 2004;304(October):457–61.
7. Freedman EG. Coordination of the Eyes and Head during Visual Orienting Edward. Exp Brain Res. 2008;190(4):369–87.
8. Tollin DJ, Populin LC, Moore JM, Ruhland JL, Yin TCT. Sound-localization performance in the cat: The effect of restraining the head. J Neurophysiol. 2005;93(3):1223–34.
9. Bracis C, Gurarie E, Van Moorter B, Goodwin RA. Memory effects on movement behavior in

animal foraging. PLoS One. 2015;10(8):1–21.

10. Knight K. Head movements give away foraging behaviour. Vol. 214, Journal of Experimental Biology. Company of Biologists Ltd; 2011. p. i–ii.

11. Berger M, Agha NS, Gail A. Wireless recording from unrestrained monkeys reveals motor goal encoding beyond immediate reach in frontoparietal cortex. Elife. 2020;9:1–29.

12. Grabenhorst F, Tsutsui KI, Kobayashi S, Schultz W. Primate prefrontal neurons signal economic risk derived from the statistics of recent reward experience. Elife. 2019;8:1–30.

13. Leon MI, Shadlen MN. Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. Neuron. 1999;24(2):415–25.

14. Tsutsui KI, Grabenhorst F, Kobayashi S, Schultz W. A dynamic code for economic object valuation in prefrontal cortex neurons. Nat Commun. 2016;7.

15. Miller EK, Erickson CA, Desimone R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. J Neurosci. 1996;16(16):5154–67.

16. Khalighinejad N, Bongioanni A, Verhagen L, Folloni D, Attali D, Aubry JF, et al. A Basal Forebrain-Cingulate Circuit in Macaques Decides It Is Time to Act. Neuron. 2020;105(2):370-384.e8.

17. Hernández-Peón R, Scherrer H, Jouvet M. Modification of electric activity in cochlear nucleus during “attention” in unanesthetized cats [9]. Vol. 123, Science. American Association for the Advancement of Science; 1956. p. 331–2.

18. Fernandez-Leon JA, Parajuli A, Franklin R, Sorenson M, Felleman DJ, Hansen BJ, et al. A wireless transmission neural interface system for unconstrained non-human primates. J Neural Eng. 2015;12(5):56005.

19. Yin M, Li H, Bull C, Borton DA, Aceros J, Larson L, et al. 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. 2013. p. 3109–14.

20. Milton R, Shahidi N, Dragoi V. Dynamic states of population activity in prefrontal cortical networks of freely-moving macaque. Nat Commun. 2020;

21. Shahidi N, Schrater P, Wright T, Pitkow X, Dragoi V. Population coding of strategic variables during foraging in freely-moving macaques. bioRxiv. 2019;811992.

22. Herrnstein RJ. Relative and absolute strength of response as a function of frequency of reinforcement. J Exp Anal Behav. 1961 Jul;4(3):267–72.

23. Chandrasekaran C, Peixoto D, Newsome WT, Shenoy K V. Laminar differences in decision-related neural activity in dorsal premotor cortex. Nat Commun. 2017;8(1).

24. Elliott R, Friston KJ, Dolan RJ. Dissociable neural responses in human reward systems. J Neurosci. 2000;20(16):6159–65.

25. Pierce DW, Cheney CD. Schedules of Reinforcement. B. F. Skinner Foundation. 2015.

26. Stevenson IH, Kording KP. How advances in neural recording affect data analysis. Nat Neurosci. 2011;14(2):139–42.

27. Bunge SA, Burrows B, Wagner AD. Prefrontal and hippocampal contributions to visual associative recognition: Interactions between cognitive control and episodic retrieval. Brain Cogn. 2004;56(2 SPEC. ISS.):141–52.

28. Pitkow X, Angelaki DE. Inference in the Brain: Statistics Flowing in Redundant Population Codes. Neuron. 2017;94(5):943–53.

29. Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, Poeppel D. Neuroscience Needs Behavior: Correcting a Reductionist Bias. Neuron. 2017;93(3):480–90.

30. Wu Z, Schrater P, Pitkow X. Inverse Rational Control: Inferring What You Think from How You Forage. 2018;

31. Calapai A, Berger M, Niessing M, Heisig K, Brockhausen R, Treue S, et al. A cage-based training, cognitive testing and enrichment system optimized for rhesus macaques in neuroscience research. Behav Res Methods. 2017;49(1):35–45.

32. Fagot J, Gullstrand J, Kemp C, Defilles C, Mekaouche M. Effects of freely accessible computerized test systems on the spontaneous behaviors and stress level of Guinea baboons (Papio papio). Am J Primatol. 2014 Jan;76(1):56–64.

33. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018;21(9):1281–9.

34. Bala PC, Eisenreich BR, Bum S, Yoo M, Hayden BY, Park HS, et al. OpenMonkeyStudio : Automated Markerless Pose Estimation in Freely Moving Macaques. Bioarxiv. 2020;1–12.

35. Bishop C.M. Pattern recognition and Machine learning. 1st ed. M. Jordan, J. Kleinberg, B. Scholkopf, editors. New York: Springer; 2006.

36. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009;10(3):515–34.

Supplementary figures

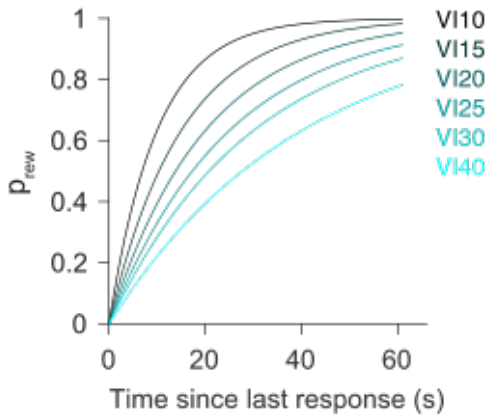

Figure S1 | The probability of reward availability as a function of the scheduled reward rate and the time since the preceding response on the same box.

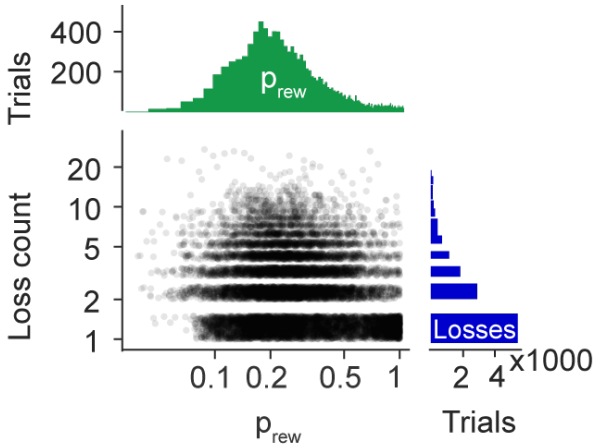

Fig S2 | Joint distribution of the *loss count* and p_{rew} for 13,269 responses, and the marginal distributions of p_{rew} and the *loss count*.

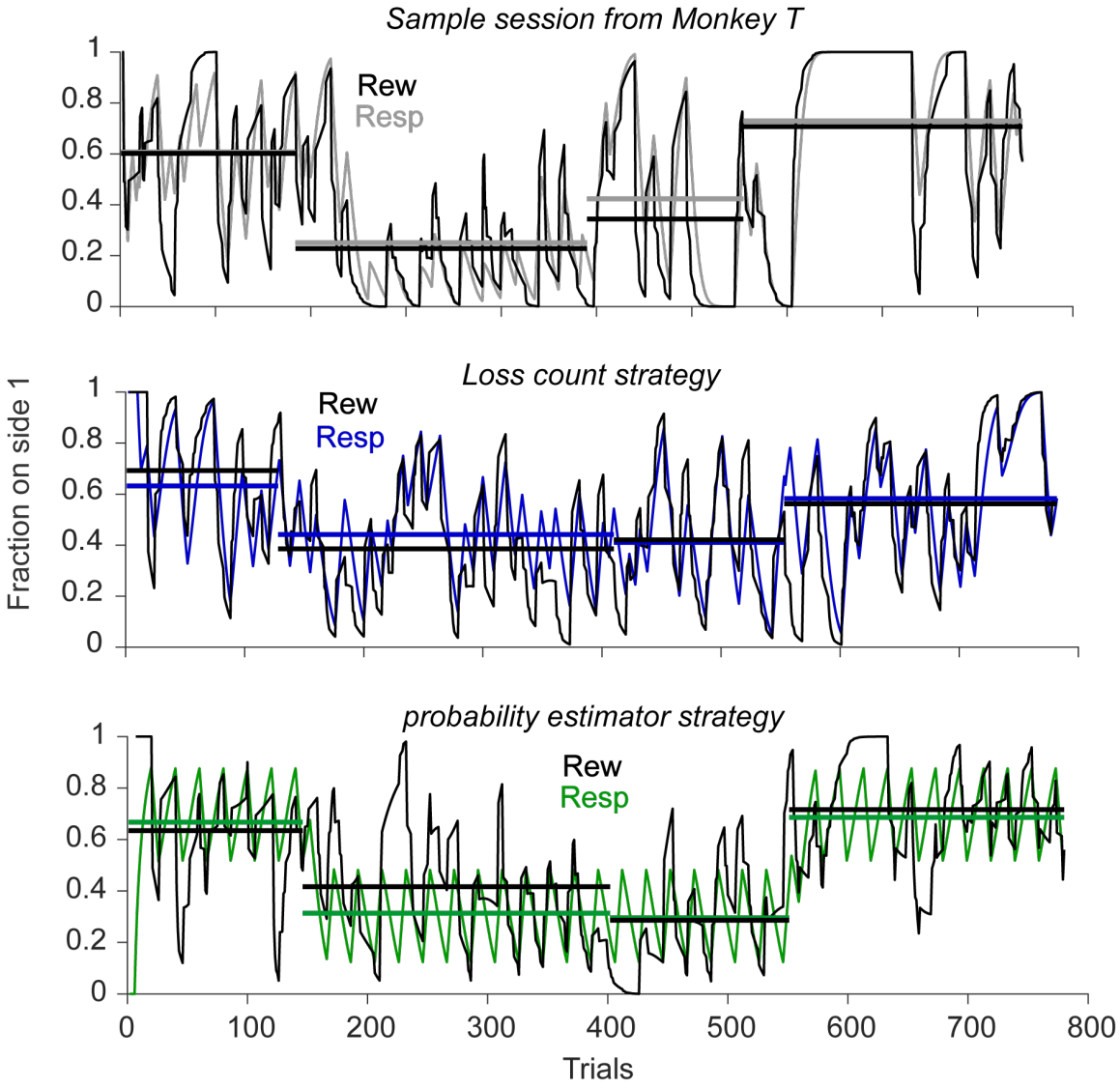

Figure S3 | Dynamic matching for a sample session of Monkey T with 3 sets of reward schedules: VI15-VI25, VI25-VI15, and VI15-VI25 again. We compared two simulated agents, one with loss counting strategy (blue) and the other with probability estimating strategy (green). The reward and response rates were calculated locally using a causal Gaussian filter (5).

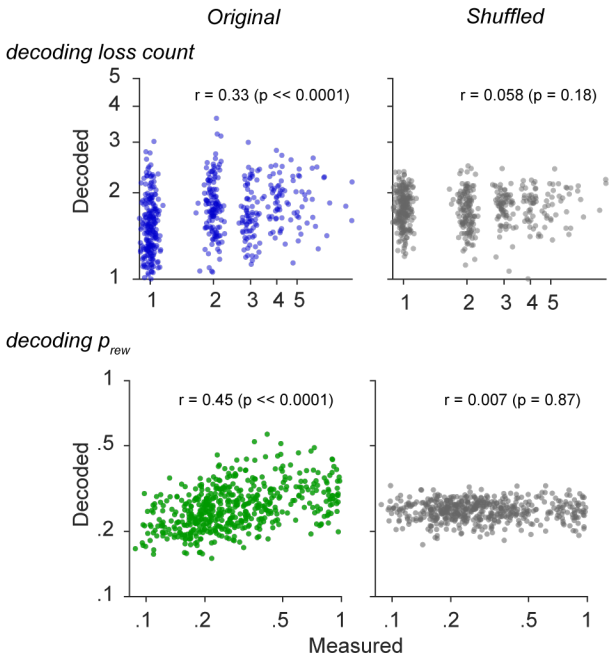

Figure S4 | decoding task variables for two sample sessions. Left: Each task variable was decoded using the pre-response activity of the entire simultaneously recorded neural population. A regression model was trained and tested as described in the methods section. The decoding performance was measured as the Pearson correlation confident between the measured and the decoded task variable. Right: As a control, decoding was repeated for shuffled labels (see methods).

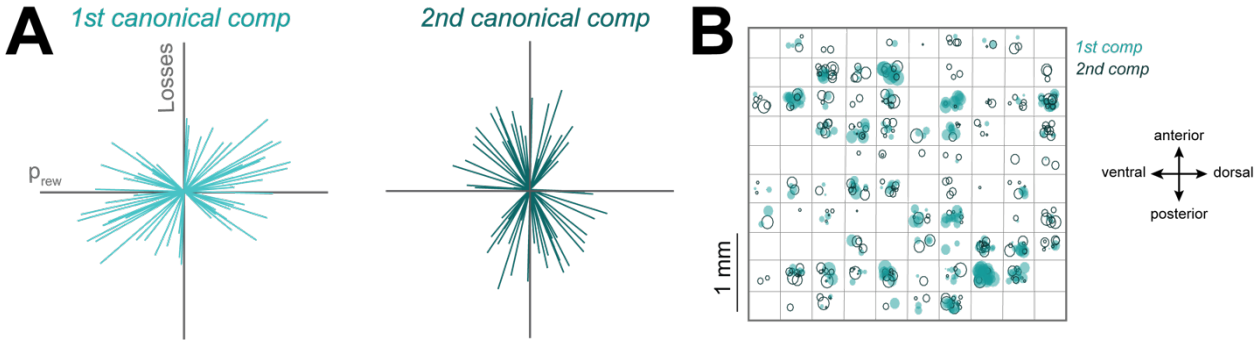

Figure S5 | (A) The placement of canonical components in the 2D space of p_{rew} and *loss count* for 33 sessions. The length of the line segments is proportional to the correlation coefficient between pairs of components in this space and neural space. **(B)** Contribution of each neuron in the first and second canonical components for one example recording session. The diameter of each circle is proportional to the weight of that neuron for the corresponding color-coded canonical component. The location of each circle shows the anatomical location of the recorded neuron on the map of the multi-electrode array.

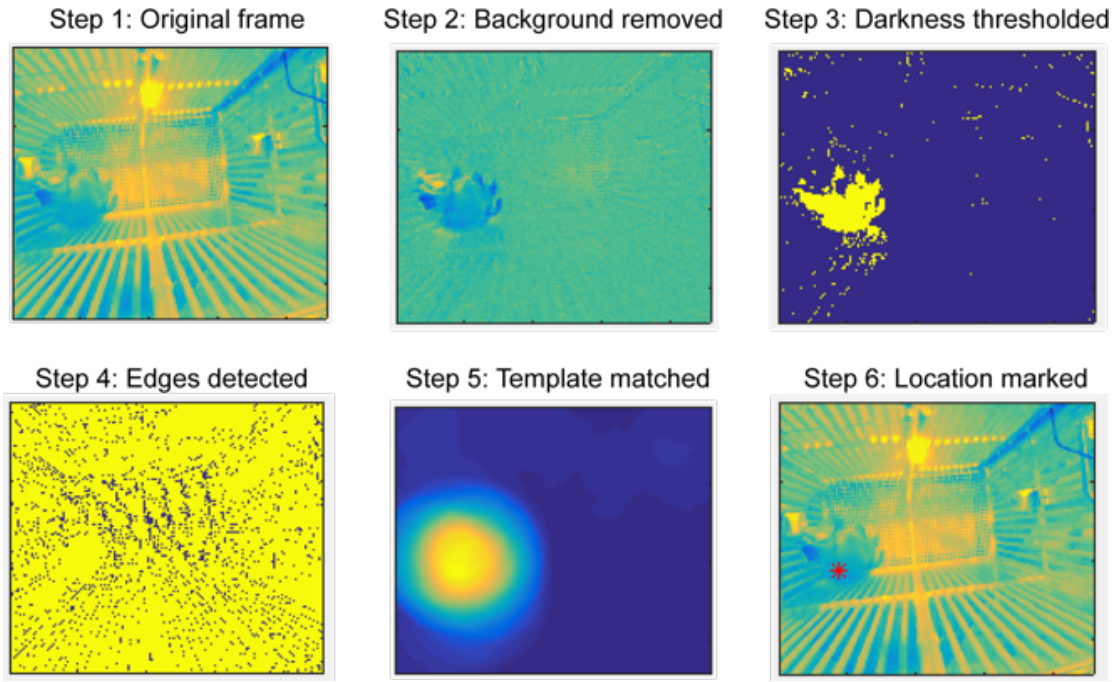

Fig S6 | Sequential processing of images from the overhead camera, used to locate the monkey. See Methods for the description of steps 1-6**.**

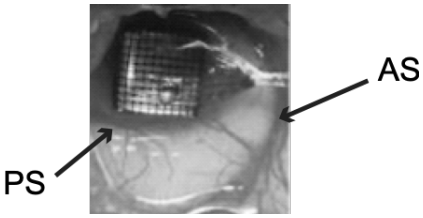

Figure S7 | The location of a 96-channel Utah array in dlPFC (area 46) on the left hemisphere of monkey G. The arcuate sulcus (AS) and principal sulcus (PS) are marked.

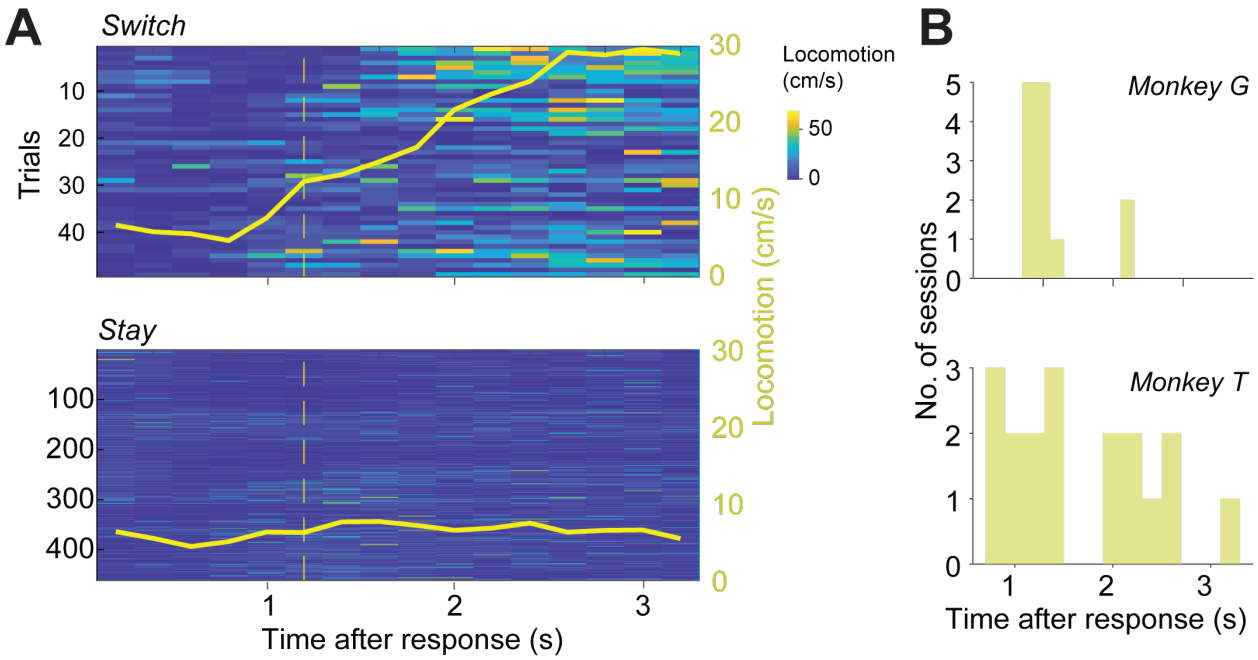

Fig S8 | Detecting the time of the switches using the locomotion data. **(A)** the magnitude of locomotion in each 200 ms time bin, separated for switch and stay trials. The trial averages are shown as solid yellow lines. The yellow dashed line shows the time in which the average magnitude of locomotion in switch trials passes the 3 SD of the magnitude of locomotion in stay trials. **(B)** Average switch time in each session.