Meta-learning is expressed through altered prefrontal cortical dynamics

Xulu Sun*[1,2], Alison E. Comrie*[1,2,3], Ari E. Kahn[4], Emily J. Monroe[1], Abhilasha Joshi[2], Jennifer A. Guidera[5], Eric L. Denovellis[2], Timothy A. Krausz[3], Jenny Zhou[6] , Paige Thompson[6] , Jose Hernandez[6] , Allison Yorita[6] , Razi Haque[6], Joshua D. Berke[7,8], Nathaniel D. Daw[4,9], Loren M. Frank[1,2,7,10]

[1] Department of Physiology and Psychiatry, University of California, San Francisco; San Francisco, CA 94158, USA
[2] Howard Hughes Medical Institute; Chevy Chase, MD 20815, USA
[3] Neuroscience Graduate Program, University of California San Francisco; San Francisco, CA 94158, USA
[4] Princeton Neuroscience Institute, Princeton University; Princeton, NJ 08544, USA
[5] Medical Scientist Training Program, University of California, San Francisco, San Francisco, CA 94143, USA
[6] Center for Micro- and Nano-Technology, Lawrence Livermore National Laboratory, Livermore, 94550, USA
[7] Kavli Institute for Fundamental Neuroscience, University of California, San Francisco; San Francisco, CA 94158, USA
[8] Department of Neurology and Department of Psychiatry and Behavioral Science, and Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA 94158, USA
[9] Department of Psychology, Princeton University; Princeton, NJ 08544, USA
[10] Lead contact
* These authors contributed equally
Correspondence: xulu.sun@ucsf.edu, comriea@janelia.hhmi.org, loren.frank@ucsf.edu

Abstract

Learning where and when rewards like food and water are available is essential for survival[1,2]. In the simplest cases where resource availability is stable, animals can learn reward contingencies by integrating outcomes across repeated samples of each option. In more natural settings, however, reward availability is governed by complex contingencies such as depletion and repletion over time. To flexibly adapt to such changing environments, optimal choices require meta-learning wherein animals learn how to learn from external feedback, eventually enabling them to infer reward contingencies from more abstract and generalizable rules rather than relying solely on recent outcomes[3,4]. The existence of meta-learning in animal behavior is well established[3,5–8], yet the neural circuits and computations that implement it remain poorly understood[9–11]. Here we investigated meta-learning using a spatial foraging task in which rats acquired a novel depletion-repletion rule, and carried out longitudinal, high-density recordings from the medial prefrontal cortex (mPFC). We show that meta-learning engages specific, systematic changes in mPFC neural dynamics that embed the learned rule and thereby alter the way the network learns option values from reward outcomes. These dynamics are based on mixed coding of task structure and value in individual mPFC neurons which, at the population level, organize into low-dimensional dynamical motifs that generalize across task conditions. As meta-learning progresses, these motifs are reshaped to instantiate both rule-guided inference of future states before outcome delivery and rule-based value updating during the outcome period. These results indicate that meta-learning

sculpts pre-existing prefrontal dynamics to support the acquisition of new reward-learning strategies.

Main

To make adaptive decisions in a changing environment, the brain must learn to incorporate external feedback into its evaluation processes and action policies. For instance, effective foraging in the wilderness requires animals to learn how food, water, or other rewards are distributed and use this knowledge to predict the likely outcomes of their choices. In stable environments, such predictions are straightforward: repeated sampling of each option eventually converges on accurate estimates of reward contingencies. However, many natural environments impose additional structure—such as resource depletion with use and recovery over time—that demands learning not only immediate reward availability but also the rules governing how that availability evolves. This higher-order learning process, known as meta-learning, allows the brain to improve its own learning algorithm by abstracting regularities across experiences[3,4,12,13]. Through meta-learning, animals can generate more accurate predictions of action values and consequently make more advantageous choices. It also enables generalizing strategies across contexts that share common rules[11,14,15].

Converging evidence from recent experimental and theoretical studies points to the prefrontal cortex (PFC) as a critical substrate for meta-learning. Empirical work shows that PFC is necessary for refining learning algorithms and stabilizing value coding over experience[10]. Complementary computational models of prefrontal recurrent networks demonstrate that these circuits can implement rapid value updating, while the mechanism of this updating process—effectively the learning rule itself—is shaped by slower meta-learning that incorporates higher-order task statistics[9,16]. Such statistics often define how latent task states evolve over time to enable prediction beyond immediate outcomes[9,17]. Further, PFC has been implicated in abstraction and generalization[18–20], consistent with the proposition that meta-learned strategies could be expressed flexibly across different task conditions or contexts[14,21]. Despite these insights, it remains unclear whether meta-learning drives systematic changes in PFC population dynamics to implement altered value computations and support rule-based inference of future task states. Moreover, whether similar dynamics are reused to subserve generalized computations across contexts is also unknown.

Here we focus on the medial PFC (mPFC), a brain region known to represent abstract task structure, encode action–outcome contingencies, and predict future states[2,13,18,19,22–25]. We test the hypothesis that meta-learning sculpts mPFC population dynamics, altering the computations that underlie task-state inference, value updating, and action selection. We used a multi-patch foraging task in which rats first learned stable reward contingencies of each patch from primarily caching reward history[26] and then, over days, acquired an abstract depletion–repletion rule that controlled reward availability across patches. We developed a statistical behavioral model both to infer trial-by-trial subjective values predictive of rat decisions and to quantify the meta-learning process underlying shifts in reward-learning strategy. This task and model, combined with stable, high-density polymer probe recordings from mPFC over many tens of sessions[27], allowed us to capture the reorganization of mPFC population dynamics as animals adapted their learning strategy to incorporate the rule.

We found that mPFC neurons simultaneously tracked task structure (e.g., goal progress, action, and foraging patch identity) and value through multiplicative coding. Building upon the single-neuron mixed-coding profiles, coordinated population activity formed dynamical motifs that instantiated value- and decision-relevant computations and generalized across task conditions.

Crucially, meta-learning reshaped these motifs to support rule-based value inference, paralleling the emergence of a new reward-learning strategy. These results point to a general computational principle whereby structured population dynamics can internalize task rules and flexibly adapt reward-learning algorithms across contexts. More broadly, the identification of recurring, low-dimensional dynamical motifs helps elucidate how the brain abstracts across experiences, a core property of intelligence[28–30].

Behavioral evidence for meta-learning in the spatial bandit task

We developed a variant of a spatial bandit task[26] in which rats (n=4) foraged for probabilistic rewards distributed across multiple patches (Fig. 1A; see Methods). The environment consisted of three Y-shaped 'foraging patches' radiating from the maze center, and each patch contained two reward ports at path ends. The multiple bifurcations and reward sites in our task exposed animals to a broader set of decision-making conditions than in standard two-arm bandit designs. This richer structure enables identification of generalized computational principles proposed in recent theoretical work[28,29].

In this task a trial was defined as the interval between successive departures from two reward ports. Each port was assigned a nominal reward probability, p(R), of 0.2, 0.5, or 0.8, with one patch offering the highest average p(R) under a given assignment. On each visit to a port, reward delivery was determined by its p(R), and consecutive visits to the same port were never rewarded. During the first phase of training (1-2 weeks), the p(R) of all ports remained stable for blocks of 60 trials and changed covertly at block transitions within each session. As shown in our previous work[26], rats learned to identify and exploit the patch with the highest reward probabilities and to adjust patch choices after the nominal p(R) changed.

To examine whether animals could go beyond purely outcome-based learning and acquire an abstract reward rule, we then introduced a second phase of the task. In this phase, the nominal p(R) was held fixed for each session while a new depletion-repletion rule altered the temporal structure of reward availability: if rats stayed within the same patch (i.e., alternated between its two ports), p(R) of the visited port was discounted by 80% on the next visit, and restored only when they switched to a different patch (Fig. 1B). Each rat typically completed five 300-trial sessions per day, yielding ~9500 total trials per animal. Nominal p(R) assignments were shuffled across sessions, allowing us to study both within-session reward learning and cross-session meta-learning of the new rule. Accordingly, all subsequent analyses focused on the second phase.

Inspection of behavioral patterns revealed that over days of training, rats adapted their foraging strategies as they gained experience with the depletion-repletion rule. In early-learning sessions, rats explored the three patches and typically stayed within a given patch until the reward was depleted, consistent with an action-selection strategy based on recent reward outcomes (Fig. 1C, top). In late-learning sessions, after the initial exploration phase, rats homed in rapidly on the high-reward patch, often also identifying the medium-reward patch. Furthermore, we saw evidence for a novel strategy: following exploitation of the high-reward patch, rats made brief visits to a lower-reward patch before returning to the high-reward one (Fig. 1C, bottom). Remarkably, they commonly switched back to the high-reward patch even when their visit(s) to the lower-reward patch were rewarded. This acquired strategy indicates the emergence of meta-learning—the moment-to-moment action policy was shaped by an internal model that, rather than relying on recent returns at each patch, forecasted impending depletion in the current patch and renewed availability elsewhere.

To quantify these reward-learning and decision-making patterns, we fit a statistical behavioral model to rats' trial-by-trial choices and reward outcomes (Fig. 1D; see Methods). The model

generated single-trial estimates of subjective port and patch values, enabling analysis of value-related computations associated with individual decisions (Fig. 1E). Specifically, it embedded a Bayesian version of recency-weighted reward learning, which is analogous to Q learning but infers reward probability based on a Beta-Bernoulli distribution. It was further augmented with a parameterized depletion rule that allowed obtained outcomes to deviate from the underlying reward probability. Accordingly, by fitting the model, we inferred a 'depletion factor' for each day that captured how strongly the depletion rule was incorporated into choice evaluation (Fig. 1F). This factor was close to one in early-learning sessions, which corresponds to standard recency-weighted learning (i.e., no depletion). It declined to levels near the true discounting rate of 0.8 in late-learning sessions (Fig 1F), suggesting that choices became increasingly aligned with the rule.

We further validated the presence of meta-learning in each rat's behavior by measuring changes in their decision tendencies and comparing these changes to simulations from the fitted behavioral model (Fig. 1G-L). Across days, rats became more likely to choose the high-p(R) patch when switching from other patches, and more likely to choose the medium-p(R) patch when leaving the high-p(R) patch (Fig. 1G, J). These shifts reflected knowledge of the fixed nominal reward probabilities in each session. In addition, rats left a patch immediately after reward more frequently (Fig. 1H, K) and exhibited shorter bout lengths in both high-p(R) and medium-p(R) patches (Fig. 1I, L), congruent with internalization of the depletion-repletion rule.

The model's ability to reproduce rats' behavioral patterns motivated us to use its trial-by-trial value estimates to identify a decision variable that predicted rats' choices. Choices were significantly influenced by both the value of the current patch and the maximal value of alternative patches (Fig. 1M, top). The opposing regression weights suggest that decisions were guided by a comparison of local and distant reward opportunities[31], a decision variable we termed the 'switch value' (Fig. 1M, bottom). Given its tight link to foraging behavior, we reasoned that the neural basis of this switch value (hereafter, 'value' for short) offers a mechanistic entry point for understanding how the brain shifts from a strategy of primarily caching recent reward history to one shaped by meta-learning of higher-order reward rules.

Multiplicative coding of task structure and value at different levels of condition generality

We then set out to understand how single mPFC neurons (n = 80-500 recorded simultaneously on each day) encoded the switch value and other relevant task-related variables. As rats traveled between reward ports (defined as 'journeys'), single-neuron activity showed three notable features: (i) temporal profiles that generalized across journeys sharing common task structures—such as goal progress and action sequence—irrespective of the specific physical location (Fig. 2A); (ii) value modulation that appeared as a multiplicative gain on task-structure coding, with different neurons encoding value at distinct journey stages and at varying levels of condition generality (Fig. 2B); and (iii) value modulation across the mPFC population that collectively spanned the entire progression of each journey. We systematically characterized these encoding properties by fitting generalized linear models (GLMs) to single-neuron activity[32] (see Methods). We compared the regression coefficients of GLMs fitted separately for each journey to assess the generality of task-structure and value encoding across task conditions.

We found that a substantial fraction of recorded mPFC neurons (30-90% per rat) showed significant modulation by value, with heterogeneous patterns of condition generality in both task-structure and value coding (Fig. 2C, D; Extended Data Fig. 1): some neurons exhibited consistent regression-coefficient profiles across all stay journeys; others showed restricted generalization, only across stay journeys involving the same action (left versus right turns) or within specific patches.

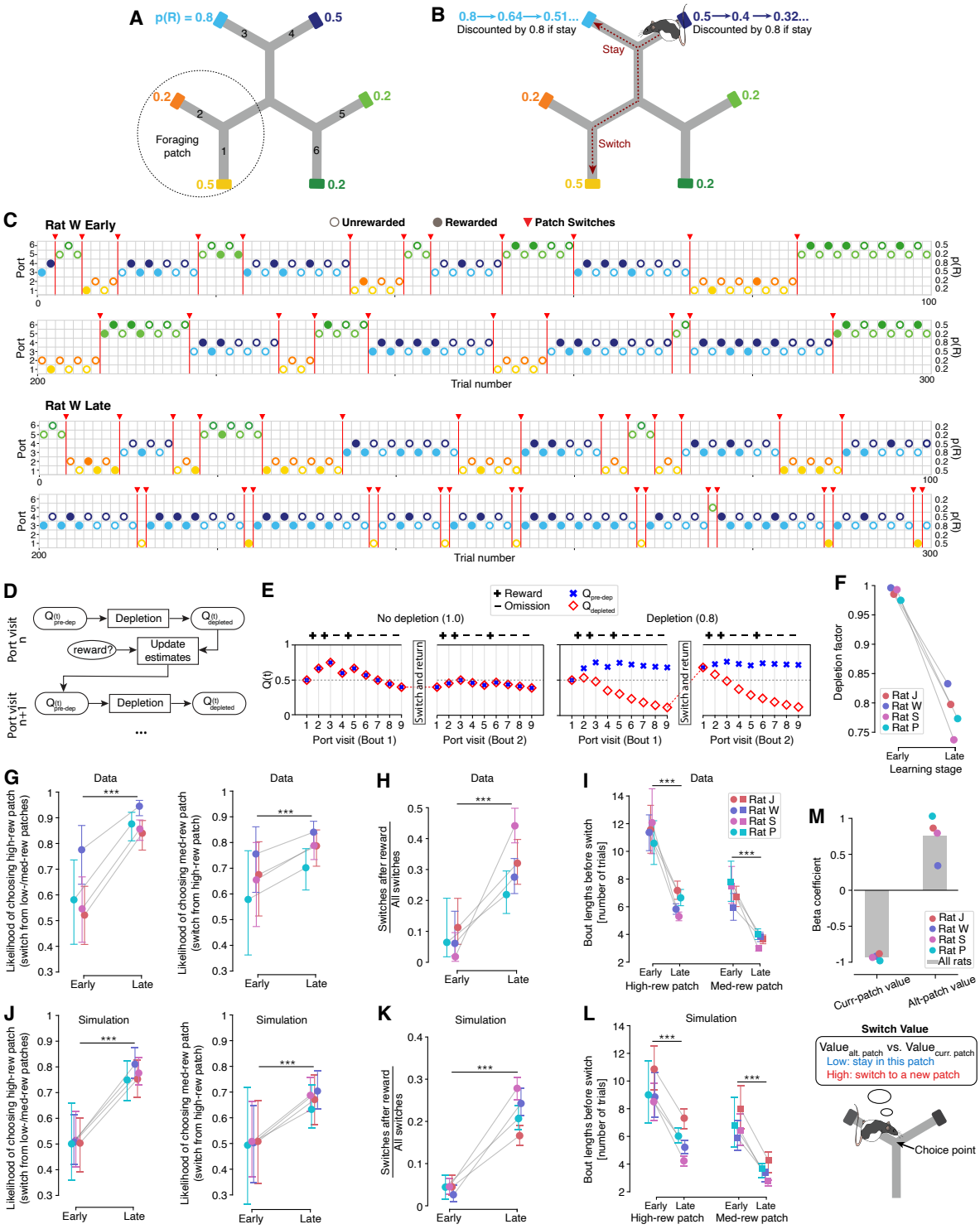


Fig. 1 | Rats expressed value-guided behavior in a spatial bandit task and learned a novel strategy. A, Maze schematic. The maze consists of three Y-shaped foraging patches radiating from a central junction, and each patch contains two reward ports at path ends. Ports deliver liquid rewards according to their assigned reward probabilities [p(R)] that vary across sessions. An example set of p(R) is shown. Ports are labeled as 1-6 and colored by their identity for figure visualization only. **B,** Task rules. Consecutive visits within the same

patch ('Stay'), deplete p(R) by a factor of 0.8 after each visit. Visiting a different patch ('Switch') restores p(R) to its nominal value. Example depletion is illustrated when rats stay in the blue patch. **C,** Example behavior from the first and last 100 trials of representative early- and late-learning sessions. Filled circles: rewarded trials; hollow circles: unrewarded trials (colored by port identity). Red triangles: patch switches. **D,** Schematic of value updating in the behavioral model with depletion. Subjective values of a visited port were first estimated as the mean of a Beta-distribution posterior updated by reward outcomes (i.e., the baseline value, $Q_{pre\text{-}dep}$), and then multiplied by a depletion factor that discounted values during repeated visits within a patch and reset upon switching patches (i.e., $Q_{depleted}$). **E,** Example Q value estimation for a given port with and without including the depletion factor. Red diamonds indicate the subjective value estimates on the current trial; blue crosses indicate belief about the baseline undepleted value; plus signs indicate reward and minus signs indicate omission. Left: without incorporating the effect of depletion, the baseline value is underestimated. Right: by quantifying the effect of depletion, the baseline value remains near its true undepleted level, while subjective value estimates decrease trial by trial as the rat stays in the same patch. *In this example, a patch is simplified as one port while in the real task, each patch consists of two ports and Q_{patch} is a weighted average of its two ports.* **F,** Depletion factor inferred from behavior decreased across days, indicating acquisition of the depletion–repletion rule. **G-I,** Altered behavioral metrics in late-learning sessions compared to early-learning sessions reflect the meta-learned new strategy. **G**, Left: Likelihood of choosing the high-reward patch when switching from other patches increased (two-proportion Z-test per animal: $P_J = 1.94 \times 10^{-7}$, $P_W = 5.10 \times 10^{-5}$, $P_S = 3.81 \times 10^{-8}$, $P_P = 5.30 \times 10^{-5}$; linear mixed effects model across all animals: $P = 1.99 \times 10^{-19}$). Right: Likelihood of choosing the medium-reward patch when switching from the high-reward patch also increased (two-proportion Z-test per animal: $P_J = 0.08$, $P_W = 0.09$, $P_S = 0.05$, $P_P = 0.14$; linear mixed effects model across animals: $P = 1.93 \times 10^{-35}$). **H**, Among all switch trials, the proportion of switches immediately after reward increased (two-proportion Z-test per animal: $P_J = 0.00044$, $P_W = 0.00068$, $P_S = 1.28 \times 10^{-9}$, $P_P = 0.024$). **I,** Bout lengths (number of stay trials before a switch) decreased for both high-reward and medium-reward patches (one-sided Wilcoxon rank-sum test: $P_J = 8.94 \times 10^{-7}$ and 1.64×10^{-14}, $P_W = 7.67 \times 10^{-10}$ and 2.72×10^{-4}, $P_S = 3.05 \times 10^{-10}$ and 3.03×10^{-18}, $P_P = 1.82 \times 10^{-5}$ and 5.39×10^{-8}). Error bars: 95% CI. **J-L,** Behavioral metrics corresponding to G-I reproduced by simulations of the fitted behavioral model (same statistical tests as in G-I: $P < 10^{-5}$ for all pairs of early- vs. late-learning comparisons). **M,** Logistic regression coefficients predicting single-trial Switch/Stay decisions from model-estimated patch values (see Methods). Values of current and alternative patches both significantly influenced rat choices, in opposite directions ($P < 0.05$ for coefficients of $Value_{curr.\ patch}$ and $Value_{alt.\ patch}$ across early and late sessions per rat). We thus defined the *switch value* as $Value_{alt.\ patch} - Value_{curr.\ patch}$. **C, F-L,** Throughout the study, early learning was defined as all run sessions on the first day of phase 2, whereas late learning comprised all run sessions from days 5–7 after the new behavioral strategy had emerged, as indicated by a pronounced reduction in the depletion factor estimated by the behavioral model.

To determine the predominant level of value-coding generality for each neuron, we then fit additional GLMs across journeys pooled by shared task structures and compared these model fits (see Methods). This approach allowed us to quantitatively classify neurons spanning a hierarchy of generalization. Consistent with the heterogeneity of cross-journey correlations of GLM coefficients, some neurons were best described by journey-specific models, some by action- or patch-dependent models, and others by progression-based models (Extended Data Fig. 1B).

Together, these results demonstrate that mPFC neurons possess diverse and hierarchically organized coding of task structure and value, ranging from broad generalization across journeys to condition-specific patterns. Importantly, the average encoding strengths across individual neurons remained stable in early and late sessions, showing no systematic changes over meta-learning (Fig. 2D). This motivated us to investigate whether meta-learning instead altered how value was represented in the population through reorganization of neural dynamics.

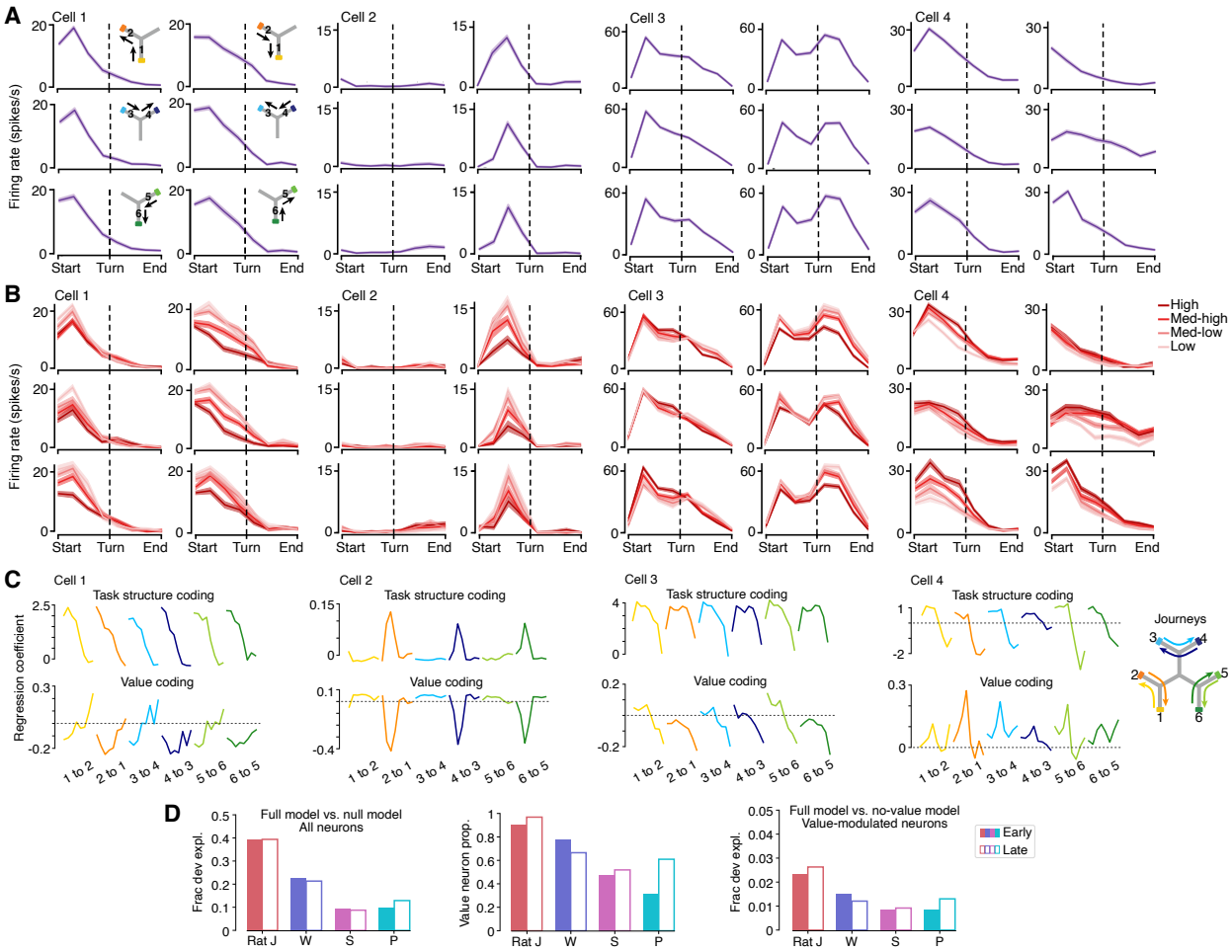


Fig. 2 | Mixed coding of task structure and value in mPFC neurons.
A, mPFC neurons were modulated by abstract task structure such as goal progress and action. Showing firing rates of four example neurons over 8 progression bins across 6 journeys. Cell 1 exhibited consistent temporal profiles across all journeys; Cell 2 was selectively active during right-turn journeys; Cell 3 was active for both left-turn and right-turn journeys but displayed action-dependent temporal profiles; Cell 4 showed broadly consistent temporal profiles across all journeys, with the range of firing rate containing specificity for patch 2 and patch 3. Inset: schematic of journeys associated with each firing profile. **B,** Value-related modulation appeared as a multiplicative gain on top of task-structure coding. Value-coding strength and sign depended on goal progress, action, patch identity, and/or journey identity. Trials of each journey were divided into four groups depending on their estimated switch value. **C,** GLM regression coefficients of 8 progression bins per journey (top) and of 8 switch value × progression terms per journey (bottom). Same example neurons as in A, B. **D,** Summary of GLM fitting results. Left: Fraction of deviance explained by the full GLM compared to the null model in early- and late-learning sessions, averaged across all neurons per rat. Middle: Proportion of value-modulated neurons in early- and late-learning sessions. Value-modulated neurons were defined as those for which the full model explained significantly higher fraction of deviance than a reduced model lacking value × progression × journey terms ($P < 0.01$). Right: Fraction of deviance explained by the value × progression × journey terms in early- and late-learning sessions, averaged across value-modulated neurons per rat. Same measurements across all days per rat are shown in Extended Data Fig. 1A to test statistical significance.

Generalized dynamical motifs for tracking, predicting, and updating values

Having observed that individual mPFC neurons encode value in conjunction with task-structure variables including goal progress, action, and patch identity, we next asked how such heterogeneous mixed codes are organized at the population level. Because meta-learning in this task required internalizing the depletion-repletion rule and applying it across conditions, we searched for dynamical motifs, i.e., recurring patterns in population activity, that could implement computations capable of tracking, predicting, and updating the switch value in different task contexts. Our aim here is to build intuition for the underlying structure of population dynamics, which we then quantify in the analyses that follow.

Given the condition generality of single-neuron coding, we first sought dynamical motifs that were shared across journeys. We grouped trials according to their proximity to a patch switch and applied principal component analysis (PCA) to neural population activity during the run period. This defined a 'navigation subspace' wherein the leading 6–7 PCs collectively explained >80% of the total variance (Extended Data Fig. 2). In this subspace, the most prominent structure was progression-tracking dynamics: neural trajectories evolved along condition-invariant loops across all journeys (Fig. 3A), as expected given the prevalence of goal-progress coding in single mPFC neurons (Fig. 2 and refs [13,18,22]). In addition, along dimensions roughly orthogonal to these loops, trajectories across trials leading up to a switch were systematically displaced in a value-related order (Fig. 3A).

Similar value-displaced loops were found at finer levels of task generality—across journeys sharing actions (left vs. right turns), across journeys in the same patch, and even within individual journeys (Fig. 3B, Extended Data Fig. 3). Thus, mPFC population activity expressed a family of common dynamical motifs that jointly represented task structure and value. These motifs are well suited to provide a unified scaffold for integrating condition-specific task variables with global value estimates to flexibly inform foraging decisions in various contexts, thus instantiating a generalized computational principle.

We next examined whether these structured population dynamics were sculpted by meta-learning in support of the rule-based strategy. Because neural variance in the navigation subspace was dominated by goal progress, we sought to extract additional dimensions that captured variance more strongly related to value. We therefore applied PCA to population activity during the pre-move time window (-0.8 to +0.2 s from trial start), a subset of the pause period when rats were stationary at reward ports and about to initiate the next movement (Extended Data Fig. 4). This yielded a 'pre-move subspace', into which we projected neural activity spanning the full pause and run periods.

Within the pre-move subspace, we identified a spiral dynamical motif characterized by a salient value-related gradient across pre-switch and switch trials (Fig. 3C, Extended Data Fig. 5). Across pre-switch trials in both early- and late-learning sessions, neural trajectories progressed along the spiral in the switch direction, paralleling the ramp-up of switch value driven by reward depletion. On switch trials, these trajectories advanced further along the spiral before movement onset. A critical feature of this motif that changed markedly with meta-learning was how neural trajectories evolved as rats traversed to a new patch, before any reward outcomes were revealed.

In late-learning sessions when rats expressed the meta-learned strategy, switch-trial trajectories 'reset' prominently away from the switch direction, typically extending well beyond stay-trial trajectories or reaching levels comparable to early-stay states in the new patch (Fig. 3C, right; Extended Data Fig. 5A-C). This anticipatory neural reset likely reflects a strong re-initialization of neural states prior to outcome delivery, consistent with inference about patch repletion. By contrast, the reset was moderate or absent in early-learning sessions (Fig. 3C, left;

Extended Data Fig. 5A-C): switch-trial trajectories either remained within the distribution of stay-trial trajectories rather than resetting further away, or showed minimal separation from switch states. The emergence of such augmented resetting suggests that task-state inference is an acquired, meta-learned computation.

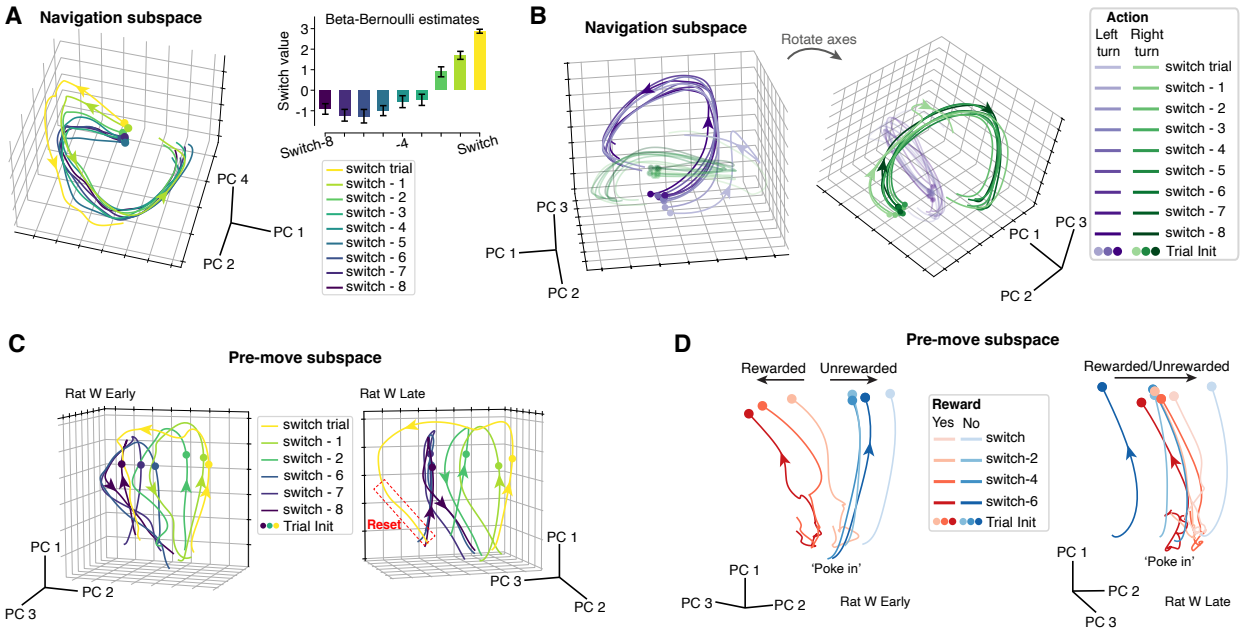


Fig. 3 | Generalized dynamical motifs related to value coding and meta-learning.
A, Run-time (0 to +3 s from trial initiation) neural trajectories across pre-switch and switch trials in the leading PC subspace constructed from run-time neural activity (one example day). Goal-progress dynamics dominated PCs 1 and 2. Approximately along an orthogonal axis (PC 4 in this example; could be PC 3 in other days), the arranging order of these trajectories reflected increasing switch value estimates. Inset: switch values computed from the Beta-Bernoulli model estimates after mean subtraction (error bars denote standard deviation). Before mean subtraction, switch values were consistently negative because the behavioral model added a patch-stay bias to the value of the current patch to capture the cost of patch switching (see Methods). **B,** Run-time neural trajectories from left-turn (purple) and right-turn (green) trials leading up to a switch in the navigation subspace (one example day). Within each action condition, trajectory displacement was systematically aligned with switch value estimates, revealing action-specific variants of the same dynamical motif. **C,** Neural trajectories in the leading PC subspace constructed from pre-move neural activity (-0.8 to +0.2 s from trial initiation) formed a spiral dynamical motif (example rat). To cover both the late value-updating period from the previous trial and the run period of the current trial, we projected neural data during the -0.8 to +3 s window (stay trials) and -0.8 to +4.5 s window (switch trials) from trial initiation. After rats entered a new patch, the switch-trial neural trajectory (yellow) did not reset as far in early-learning sessions (left) as in late-learning sessions (right). The red rectangle marks time spent in the final segment of a switch path and highlights the reset trajectory. **D,** Neural trajectories in the pre-move subspace on trials following reward versus omission (example rat). To cover the value-updating period of the previous trial before initiation of the current trial, we projected neural data during the median pause time (-5 to 0 s from trial initiation post reward and -0.8 to 0 s from trial initiation post omission). Note that among the early-session trajectories (red, left), there was no switch trial right after reward. 'Poke in' denotes arrival at the reward port. **A-D,** Circles: trial initiation; arrowheads: direction of neural trajectory evolution. Trajectories are visualized along PC axes and viewpoints chosen to highlight value-related gradients.

In addition to pre-outcome state inference, we asked how the spiral motif mediates value updating during the outcome period and whether the updating mechanism, a crucial differentiator between recency-weighted and rule-based strategies[33], changes with meta-learning. In early-learning sessions, spiral trajectories exhibited clear outcome-dependent divergence during the pause period. Specifically, they unfolded in opposite directions following reward versus omission (Fig. 3D, left). Strikingly, in late-learning sessions, neural trajectories in value-related dimensions evolved toward the switch direction across pre-switch trials regardless of preceding reward outcomes (Fig. 3D, right). This pattern is consistent with knowledge of depletion overriding the initial outcome-driven value updating.

Value-coding axes capture representational features of the dynamical motif

The PCA-revealed motifs pointed to systematic changes with meta-learning in the dynamical structure supporting value computation. To directly quantify this structure, we constructed cross-validated decoders of the switch value from mPFC population activity. Given that PCA projections suggested generalized value representations across task conditions that were consistently intertwined with goal progress, we first targeted the most general case: the conjunctive coding of goal progress and value embedded in the spiral motif spanning both run and pause periods. To this end, we pooled all stay trials and discretized each journey into eleven progression bins, including eight during run time and three during the pause (outcome) period. We then trained a separate value decoder for each bin using LASSO regression (see Methods).

These decoders explained a substantial fraction of the switch value variance (15-60%) throughout progression, with peak performance in pre-choice bins where value was most relevant to upcoming decisions (Fig. 4A, Extended Data Fig. 6A). Further, during the pause period, decoding performance was higher for the updated value of the next trial than for the value of the current trial (Fig. 4A, Extended Data Fig. 6A), pointing to an active value-updating mechanism that bridged value representations in successive decision episodes (Extended Data Fig. 6B).

Having established robust decoding performance, we next examined how neural value representations evolved across trials and related to decision-making. We therefore averaged value readouts across pre-choice bins, which were overall stable, to obtain a single pre-choice value per trial. We observed that, in both early and late sessions, decoded switch value ramped up across pre-switch trials (Fig. 4B), yielding significantly higher values on switch than on stay trials (Fig. 4C). This graded increase mirrored the arrangement of neural trajectories in the spiral motif. Moreover, decoded switch value in pre-choice bins accurately predicted single-trial Switch/Stay decisions in all rats, highlighting the behavioral relevance of the neural value representations (Fig. 4D, Extended Data Fig. 6C). Notably, the value-decoding axes during both run-initiation and pause-time bins aligned with directions in the PC subspace along which neural trajectories were displaced according to value (Fig. 4E). Together, our decoders provide an explicit low-dimensional readout of value computations implemented in the spiral dynamical motifs.

Finally, we evaluated whether such value computations were expressed consistently across multiple levels of condition generality, e.g., for journeys within the same patch, for left- versus right-turn journeys, and for each individual journey. We fit decoders at each of these generality levels and they all achieved reliable performance (Extended Data Fig. 6D; see Methods). This decoding robustness aligns with the coexistence of value-related dynamical motifs across task conditions and further corroborates the generalizability of the underlying value computation principles.

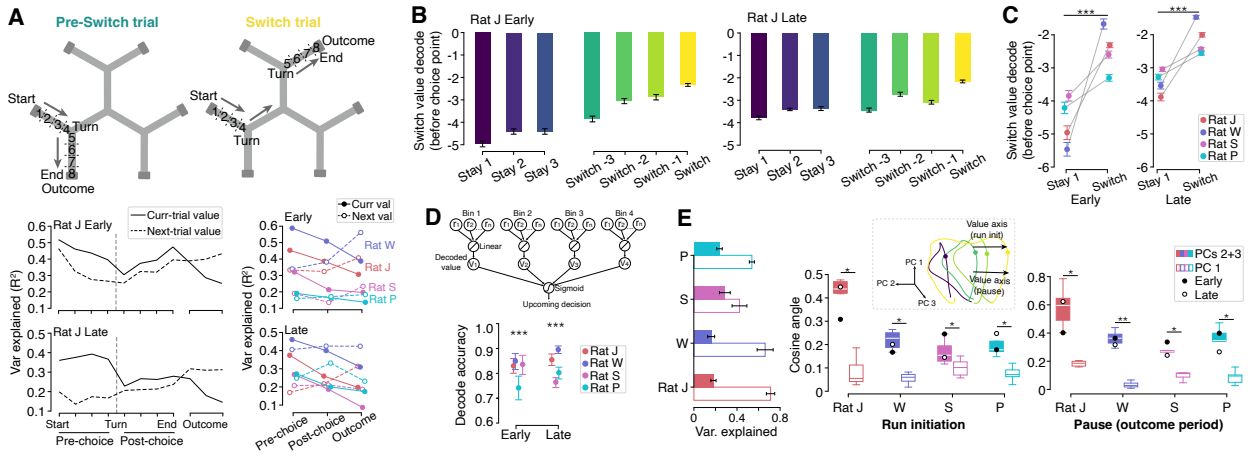


Fig. 4 | A linear value decoder approximates value-coding and value-updating dimensions.
A, Decoder design and performance. Top: Schematic showing progression-bin division of each trial used in value decoding. Switch trials were always held out as test data. Bottom left: Variance of switch value (estimated by the behavioral model) explained by the neural decoder in each progression bin (example rat, early- vs. late-learning sessions). Bottom right: Decoder performance averaged across pre-choice (bins 1-4), post-choice (bins 5-8), and outcome (bins 9-11) bins (early vs. late sessions of all rats). During the pre-choice period, the decoder explained more variance of the current-trial value than of the next-trial value; during the outcome period, this pattern reversed. When decoding current-trial value, the performance was higher in pre-choice bins. Statistical significance of these trends across all days and animals is tested in Extended Data Fig. 6A. **B,** Decoded switch value during the pre-choice period across early stay trials (Stay 1-3), late stay trials preceding a switch (Switch-3 to Switch-1), and the switch trial, shown for early- and late-learning sessions of an example rat. **C,** Decoded switch value increased significantly from the first trial ('Stay 1') to the final trial ('Switch') initiated in a given patch in both early and late sessions (linear mixed effects models testing random slopes per animal: $P_J = 1.23 \times 10^{-72}$, $P_W = 2.73 \times 10^{-75}$, $P_S = 1.78 \times 10^{-27}$, $P_P = 9.09 \times 10^{-17}$ in early sessions; $P_J = 3.06 \times 10^{-105}$, $P_W = 2.29 \times 10^{-181}$, $P_S = 1.06 \times 10^{-52}$, $P_P = 8.42 \times 10^{-37}$ in late sessions), consistent with the displacement of neural trajectories in the PC subspace (Fig. 3A, C). Error bars: SEM across trials. **D,** Decode decisions from neural value representations. Top: schematic showing the construction of value and decision decoders. r_1-r_n: neural population activity. V_1-V_4: decoded switch value in progression bins 1-4. Bottom: decoding accuracy of Switch/Stay decisions in early and late sessions (bootstrap test for accuracy > 0.5: P = 0 in all rats). To avoid accuracy inflation due to imbalanced class counts, we measured the decoding performance for Switch and Stay decisions separately (Extended Data Fig. 6C) and took the mean of them. Error bars: 95% CI from bootstrapping. **E,** Alignment between value-decoding axes and PC dimensions. Left: Across all days per rat, PC 1 explained larger variance of the pre-move neural population activity than PCs 2+3. Right: Cosine similarity between PCs 1-3 of the pre-move subspace and the value-decoding axis of the first progression bin ('run init') or the outcome-period bin ('pause'). Because spiral trajectories showed pronounced value-related displacement along PC 2, PC 3, or their combination (varying across days and animals), we computed the cosine angle between the value-decoding axis and the PCs 2+3 plane (filled boxes). Boxes include all days of each animal to test statistical significance; early (filled circles) and late (hollow circles) days are highlighted to demonstrate stability of this alignment. By contrast, alignment with PC 1 was significantly weaker (empty boxes), supporting the idea that the correlation between the value-decoding axis and PCs 2+3 was not due to chance (one-sided paired Wilcoxon signed-rank test: $P_J = 0.016$, $P_W = 0.016$, $P_S = 0.031$, $P_P = 0.031$ during run initiation; $P_J = 0.016$, $P_W = 0.0078$, $P_S = 0.016$, $P_P = 0.031$ during pause). Inset: value-decoding axes projected into the pre-move PC subspace in a representative day.

Changes in the value-coding dynamics over meta-learning

The neural value decoder provides a direct means to quantify the reorganization of mPFC population dynamics associated with meta-learning, as indicated by the reshaped spiral trajectories in late sessions. We therefore examined longitudinal changes in neural value representations that accompanied the internalization of the depletion-repletion rule. Motivated by the pronounced resetting of spiral trajectories in late-learning sessions, we asked whether neural value representations exhibited similar resets. To do so, we compared the decoded switch value at patch entry before outcome delivery ('switch-in value') with that on the first stay trial in the same patch ('first-stay value'), restricting analyses to rewarded switch-in trials to ensure comparability across early and late learning (Fig. 5A, B). Stronger neural resetting corresponds to switch-in values that lie farther below the subsequent first-stay values and measures a relative re-initialization of the value representation.

We observed that greater neural value resetting emerged with meta-learning. In early sessions, decoded switch value showed little or no reset upon entry into a new patch but decreased after rewards (Fig. 5A, C; Extended Data Fig. 7A). This implies that, early in training, re-initialization of neural value representations was primarily driven by reward receipt instead of anticipatory adjustments at the time of switching. By contrast, in late sessions, the switch-in value lay substantially below the first-stay value (Fig. 5B, C; Extended Data Fig. 7B). This pattern reflects an initialization of value for the newly entered patch that is consistent with inference about its repletion before entry and its subsequent depletion, rather than updating from direct reward experience (Fig. 5C). Because switch trials were always held out during decoder training, these neural value resets cannot be attributed to structure imposed by the behavioral model.

Besides pre-outcome inference, the pre-move dynamical motifs also suggested that meta-learning altered how the mPFC population responded to rewards and omissions (Fig. 3D). We therefore examined the evolution of neural value representations following rewarded and unrewarded switch trials. We quantified the neural value update as the change in decoded switch value from the end of a switch trial to the start of the subsequent stay and then compared updates after opposite outcomes (Fig. 5D). We found that, during early learning, decoded switch value was updated more positively after omission than after reward, consistent with outcome-based updating. By contrast, opposite reward outcomes no longer produced opposing updates in late-learning sessions (Fig. 5D, E; Extended Data Fig. 7C, D). This shift signals a transition toward rule-based value updating.

The decoupling of value updating from immediate reward outcomes was most clearly expressed in an extreme case of the meta-learned strategy: in several late sessions, rats visited lower-reward patches for a single trial before returning to the replenished high-reward patch, even when that brief visit was rewarded (Fig. 1C, bottom). This strategy exploited repletion and would generally not be expected to arise under baseline recency-weighted reward learning. These trials thus presented a unique opportunity to distinguish purely reward-driven from rule-based neural updating mechanisms. We investigated single-trial neural value representations during these single-visit episodes.

We found that changes in the decoded switch value from switching in to switching out were similar for rewarded and unrewarded single visits (Fig. 5F). In both cases, the updated values at the start of the subsequent switch-out trials lay closer to the switch boundary (Fig. 5G, left; Extended Data Fig. 7E). Further, these switch-out values were consistently higher than the values decoded for stay trials in the same patch (i.e., when rats remained in the patch after switching in), regardless of the single-visit reward outcomes (Fig. 5G, right; Extended Data Fig. 7E). These results, together with the neural value resets, demonstrate that during meta-learning (here, internalizing the depletion-

repletion rule), frontal network dynamics undergo systematic reorganization associated with pre-outcome value inference and rule-guided value updating, thereby aligning decisions with the latent structure of the environment.

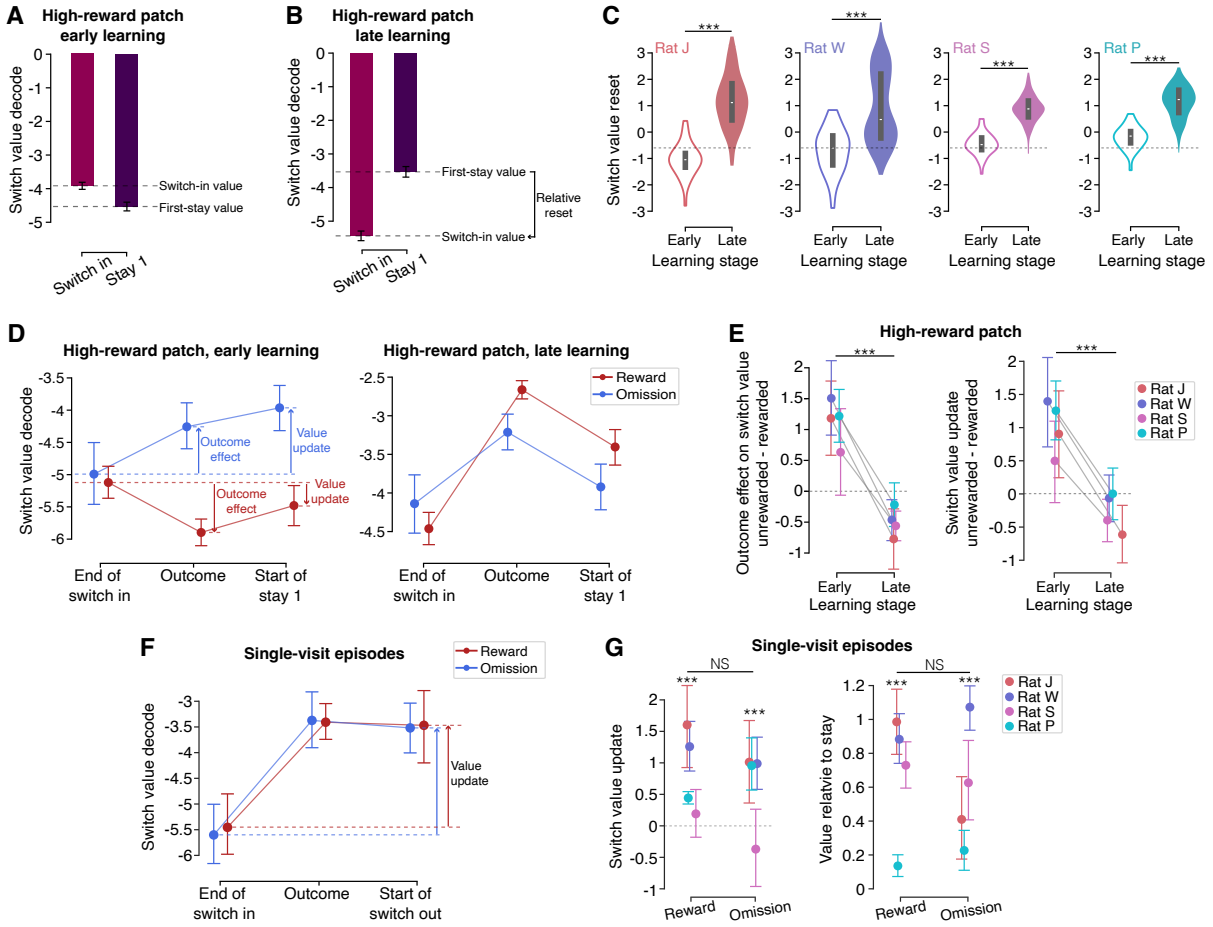


Fig. 5 | Changes in neural value representations reflect the meta-learned strategy.
A, B, Decoded switch value during the post-choice period (mean of bins 5-8, see schematic in Fig 4A) across the switch-in trial and its subsequent stay in the high-reward patch (example rat). Error bars: SEM. **A,** On an example early-learning day, decoded switch value did not re-initiate when the rat entered a new patch (switch-in value > first-stay value). **B,** On an example late-learning day, decoded switch value reset substantially when rats entered a new patch (switch-in value < first-stay value). **C,** Neural value reset, quantified as *First-stay value – Switch-in value*, increased significantly in late-learning days for all rats (one-sided Wilcoxon rank-sum test: $P_J = 1.40 \times 10^{-19}$, $P_W = 1.51 \times 10^{-13}$, $P_S = 1.31 \times 10^{-23}$, $P_P = 3.76 \times 10^{-17}$). **D,** Reward outcomes (reward vs. omission) updated neural value representations in opposite directions in early-learning sessions (left) but not in late-learning sessions (right). One example rat. We define (i) 'outcome effect' as decoded value during outcome period minus that of switch-in bin 8, and (ii) 'value update' as decoded value of first-stay bin 1 minus that of switch-in bin 8. **E,** Left: In all rats, the difference between outcome effects following unrewarded versus rewarded trials decreased significantly in late-learning sessions (linear mixed effects models testing random slopes per animal: $P_J = 1.70 \times 10^{-5}$, $P_W = 3.62 \times 10^{-7}$, $P_S = 2.02 \times 10^{-4}$, $P_P = 1.17 \times 10^{-5}$; fixed slope across all animals: 2.88×10^{-19}). Right: A similar reduction was observed for the difference in value updates (linear mixed effects models testing random slopes per animal: $P_J = 4.87 \times 10^{-4}$, $P_W = 7.84 \times 10^{-4}$, $P_S = 0.011$, $P_P = 9.12 \times 10^{-4}$; fixed slope across all animals: 1.59×10^{-9}). Error bars: 95% CI. **F,** Decoded switch value following single visits increased

regardless of whether the visit was rewarded or not (example rat). Switch value update is defined as in D except that the first-stay trial is now the switch-out trial. **G,** Neural signatures of meta-learning during single-visit episodes in late learning. Left: Switch value updates were significantly greater than 0 after both rewarded and unrewarded single visits (one-sided Wilcoxon signed-rank test across rats: $P_{rewarded} = 1.75 \times 10^{-13}$, $P_{omission} = 2.56 \times 10^{-6}$), with no significant difference between them (linear mixed effects model testing fixed slope: P = 0.24). Right: Decoded switch value on switch trials following both rewarded and unrewarded single visits was significantly greater than that on stay trials in the same patch (one-sided Wilcoxon signed-rank test across rats: $P_{rewarded} = 1.98 \times 10^{-23}$, $P_{omission} = 3.76 \times 10^{-18}$), with no significant difference between reward conditions (linear mixed effects model testing fixed slope: P = 0.72). A total of 147 rewarded and 115 unrewarded single-visit trials contributed to the analysis across all rats. Error bars: 95% CI.

Discussion

Learning requires updating internal representations based on experience, and meta-learning refines these updating processes to better match the statistical structure of the environment. Our work provides a direct window into how prefrontal cortical dynamics and computations are modified during meta-learning. We developed a spatial foraging task in which animals transitioned from learning stable reward contingencies to incorporating a higher-order depletion-repletion rule. We investigated value representations in both single mPFC neurons and population activity. Value coding in individual neurons was mixed with task-structure coding typically through gain modulation, which, at the population level, was organized into dynamical motifs that tracked goal progress, option values, and other abstract task features. These motifs generalized across task conditions, suggesting that mPFC population dynamics implement value computations in a condition-flexible manner. Notably, as animals meta-learned the depletion–repletion rule, value representations in the neural population changed markedly: neural dynamics came to express rule-guided inference of pre-outcome value and rule-based value updating, overriding purely outcome-driven computations. These findings indicate that meta-learning reorganizes mPFC dynamics underlying value tracking, updating, and prediction in support of more adaptive decision-making in complex, changing environments.

Meta-learning has been reported across a wide range of behavioral paradigms[3,5,6,11], and converging evidence implicates PFC in acquiring abstract task rules and refining learning strategies over experience[9,10,34,35]. At the single-neuron level, prior studies have shown PFC coding properties that reflect the learning of abstract rules[36,37]. More recently, population-level analyses have revealed that as animals become proficient at reward learning in a two-arm bandit task, value representations in orbitofrontal cortex (OFC) converge across days toward a stable coding axis[10]. Building on these insights, our results present several primary advances:

First, we identified a spiral dynamical motif and associated single-neuron mixed selectivity that conjunctively represent task structure and value. This motif resembles value-coding cylindrical dynamics observed in retrosplenial cortex during model-free reinforcement learning[38], suggesting that such motifs may constitute a broader class of population dynamics used across brain regions. It is worth noting that the retrosplenial motif has been attributed primarily to persistent coding of reward history, whereas the motif described here arises from mixed coding of value and task structure. Despite this distinction, the common expression of similar motifs points to a population mechanism for generalization, in which dynamical structures are reused to instantiate equivalent computations in different contexts. Critically, we found that meta-learning selectively reshaped population activity along value-coding dimensions while preserving accurate population-level representations of other variables, including goal progress, action, and patch identity. This specific

reorganization illustrates how meta-learning can flexibly adapt reward-learning algorithms without disrupting the broader representational scaffold, reconciling robustness with flexibility in PFC computation.

Second, because our task incorporated a depletion-repletion rule, animals learned to fundamentally revise their expectations about future reward availability rather than relying solely on recent outcomes. Correspondingly, neural representations of switch value re-initialized upon entry into a new patch, such that switch-in values lay well below values during the subsequent stay as animals learned that a previously depleted patch would be replenished. This relative resetting occurred before reward outcomes were revealed, indicating that mPFC dynamics developed inferential, rule-based value estimates beyond merely caching reward history. While previous studies have established a critical role for mPFC in rule-based prediction[24,39], our findings extend them by showing how prospective representations emerge through systematic, meta-learning-dependent changes in population activity.

Third, meta-learning also altered how reward outcomes influenced both behavioral policies and neural value updating. Behaviorally, animals increasingly left lower-reward patches even after a rewarded visit, marking a departure from a win-stay strategy. This shift was mirrored in mPFC population activity: early in learning, rewards consistently lowered neural representations of switch value, whereas later in learning the same outcome exerted little or no effect once the system had incorporated the depletion-repletion rule. Importantly, such reduced sensitivity to immediate outcomes did not reflect a loss of reward information. Instead, animals were still able to use reward history to estimate the repleted value of each patch and distinguish high-reward from lower-reward patches. This selective modification of outcome influence underscores the precision with which meta-learning reshapes value-updating computations.

Beyond these experimental advances, our work bears directly on a key hypothesis from computational studies: that neural network dynamics and synaptic plasticity operate as distinct learning mechanisms at complementary timescales, with the former implementing fast reward-learning algorithms and the latter mediating slower meta-learning processes that coach those algorithms[9–11]. Although synaptic mechanisms were not directly probed, our results provide empirical support for this framework by demonstrating that meta-learning sculpts mPFC dynamics underlying value- and decision-relevant computations during the emergence of a new learning strategy. In particular, the dynamical motif we identified in mPFC population activity integrates instantaneous reward learning with internalized task knowledge, establishing a direct link between evolving neural computations and behavioral strategy transitions. Such reorganization of neural dynamics is well suited to endow decision-making with greater flexibility and foresight—it allows animals to anticipate future states and select actions based on an internal model of task rules rather than assuming stable reward probabilities that can be estimated by integrating recent outcomes.

Finally, our results place mPFC dynamics within a broader framework of meta-learning that involves distributed brain regions. Other PFC subregions, including OFC[10,25,40,41], as well as subcortical regions such as the basal ganglia, have also been implicated in reward learning and meta-learning[42–45], yet how their contributions are differentiated and coordinated remains an open question[9,45]. Future work that simultaneously tracks these brain regions and their neuromodulatory inputs will be crucial for elucidating how they jointly regulate the acquisition, updating, and deployment of distinct reward-learning algorithms.

Methods

Animals

All animal procedures were approved by the Institutional Animal Care and use Committee (IACUC) at University of California, San Francisco. Five adult male Long-Evans rats (Charles River Laboratories, 450-650g) were each pair housed in a temperature- and humidity-controlled environment on a 12-hour light-dark cycle with lights on from 6 AM to 6 PM. Rats had *ad libitum* access to standard rat chow and water. Prior to behavioral training, rats were transitioned to single housing and food restricted to 85% of their baseline weight. Note that one animal was excluded from analyses because of a different training regime and a poor fit of the behavioral model, making it difficult to make clear statements about value coding comparable to other animals.

Behavioral pre-training

All behavioral tasks were controlled via custom code written in Python and Statescript in combination with an Environmental Control Unit (SpikeGadgets). The pre-training procedure was the same as described in our previous work[26]. Briefly, rats learned to run back and forth on a linear track for liquid food (sweetened evaporated milk) dispensed by reward ports at each end of the track. Upon entry of a port, 100 μL reward was delivered by a syringe pump immediately, gated by an infrared beam that rats learned to poke their noses into. The pre-training period was terminated when rats were able to gain at least 30 rewards in a 15-minute session. Pre-training took place in an environment fully separate from the spatial foraging task environment. Rats were subsequently returned to *ad libitum* food access prior to surgical implantation.

Neural implant

Custom hybrid implants contained 24 independently movable tetrodes (California Fine Wire and Kanthal) and a 3D printed tetrode drive body (PolyJet HD, Stratasys) to target the hippocampus in each hemisphere. Details about the tetrode system can be found in our previous work[26]. Implants also contained a custom headstage (SpikeGadgets) coupled with a stacking set of up to four 128-channel polymer probes[27] (Lawrence Livermore National Labs) to target the frontal cortex in each rat. The hybrid implant was protected in a 3D-printed enclosure. Drive and probes were sterilized before surgical implantation.

Both tetrodes and probes were implanted stereotaxically under sterile conditions. In anesthetized rats, cannulae were implanted above dorsal hippocampus and polymer probes were implanted in mPFC and OFC. mPFC probes were implanted with shanks (four shanks per probe) parallel to midline (+/-2.2-3.2 AP, +/-0.5-1.2 ML and -3.2-4.3 DV relative to Bregma; 0-11 degrees lateral tilt in coronal plane). Only mPFC data have been analyzed in this study. A stainless-steel ground screw was implanted over the cerebellum to serve as a global reference. Titanium screws were placed in the skull to strengthen the implant, and secured with Metabond (Parkell) and dental cement (Henry Schein).

After full recovery from surgery, rats were food deprived to 85% of their baseline weight and reintroduced to the linear track (described above), which refamiliarized them with obtaining reward on a track and prepared them for running with the implant. Two rats had additional prior experience on a fork maze[22]. This post-surgery re-training procedure ensured that rats were motivated enough to begin the spatial forging task.

Spatial foraging task with reward depletion and repletion

Before performing the main task in this study, all rats had completed 1-2 weeks of training on a no-depletion version of the spatial foraging task[26] (phase 1). In both versions of the task, animals navigated in a maze containing three radially arranged Y-shaped foraging 'patches' (Fig. 1A). Each patch consisted of a central path that bifurcated into two subpaths and terminated in reward ports, resulting in two ports per patch and six ports in total. Paths and subpaths ('track segments') measured 6.5 cm in width and 53 cm in length. All central paths intersected at 120 degrees, such that the track had three-fold symmetry. The maze was made of black acrylic (TAP Plastics) with 3-cm-high walls that allowed rats to see distal spatial cues.

Each experimental day began with transferring the rat from its home cage to a rest box located in the same room as the main task maze. The first run session started after a 20-minute baseline rest session. In the first training phase (no-depletion task), rats performed seven to eight ~20-min run sessions on each day (180 trials/session over three blocks). We then introduced the reward depletion-repletion rule. In the second phase, rats typically completed five ~30-min run sessions per day (~300 trials/session). In both phases, run sessions were interleaved with rest sessions of at least 30 minutes each. Rest sessions helped to maintain stable motivation by providing breaks throughout each day. For the depletion-repletion task, behavioral and neural data were collected over 1-2 weeks per rat. Throughout this study, early learning was defined as all run sessions on the first day of phase 2 prior to substantial rule acquisition, whereas late learning comprised all run sessions from days 5–7 after the new behavioral strategy had emerged, as indicated by a pronounced reduction in the depletion factor estimated by the behavioral model (see **Behavioral Beta-Bernoulli model** below).

During each run session, a rat was first placed at the track center facing a consistent direction, and then navigated freely in the environment under fully automated task control by Trodes software. Nominal reward probabilities p(R) of 0.2, 0.5, or 0.8 were assigned to each of the six reward ports (Fig. 1A). The two ports within a patch could share or differ in p(R), but the highest possible combination of p(R) was 0.5 and 0.8 (never 0.8 and 0.8). A trial was defined as the interval from exiting one reward port to exiting a different port (consecutive visits at the same port were never rewarded). Therefore, each trial consisted of (i) a run period between two distinct ports and (ii) an outcome period at the chosen port, during which the rat either received a 100 μL reward or experienced an omission. In phase 1 (no-depletion task), p(R) of each port remained stable within blocks of 60 trials and changed covertly at block transitions during a run session. During phase 2 (depletion-repletion task), each run session contained one fixed set of p(R) that was changed and counterbalanced across sessions. Additionally, p(R) was dynamically modulated by rats' choices: if a rat alternated between the two ports within a patch on consecutive trials, p(R) of the visited port was discounted by 80% on the next visit and continued to deplete with repeated alternation; p(R) reset to its nominal value only when the rat switched to a different patch (Fig. 1B).

All sessions took place in a dimly lit room (2.4 m by 2.9 m) with black distal spatial cues on white walls. A ceiling-mounted overhead camera (Allied Vision) centered over the maze recorded rat behavior at 30 frames/second and was synchronized via Precision Time Protocol with all other signals recorded on Trodes software (SpikeGadgets).

<u>Histology</u>

At the conclusion of behavioral experiments, rats were anesthetized with isoflurane, and small electrolytic lesions were made. One day later, animals were anesthetized and transcardially

perfused with 4% paraformaldehyde. Brains were fixed in situ overnight, after which tetrodes were retracted from the tissue and brains were extracted from the skull. Brain fixation continued for two days. After cryoprotection in 30% sucrose for 5-7 days, brains were blocked and sectioned with a cryostat into 50-100 μm sections. mPFC sections were labeled with NeuroTrace 435/455 nuclear counterstain and glial fibrillary acidic protein GFAP (mouse anti-GFAP with A594 secondary antibody) to visualize probe implant sites (Extended Data Fig. 8).

Data processing and analyses

All data processing and analyses were carried out in Python and Julia within the Spyglass pipeline to support reproducibility and data sharing[26,46].

Spike sorting and curation. We performed spike sorting using MountainSort4, an automated spike sorting algorithm[22,27]. For 128 channel polymer probe recordings in PFC, all channels were referenced against a channel on the same probe with no spikes and bandpass filtered between 300 and 6000 KHz. Artifacts were identified and removed when voltage exceeded 1500 mV on over three quarters of channels of a probe shank. Spiking units were identified in neighborhoods of channels within 115 microns of each other. All spike sorting was applied to concatenated run and sleep sessions within a day. After spike sorting, an automated round of curation was performed on the units identified by MountainSort4 using metrics calculated in SpikeInterface[47]. Units with metrics exceeding any of the following thresholds were rejected: a nearest neighbor noise overlap of 0.03, an inter-spike interval violation ratio of 0.0025, and a time shift in the waveform peak of two samples[22]. Then manual curation was performed on the remaining units to further improve the identification of single units and removal of artifacts. This process involved visual inspection to achieve two major objectives. First, different units whose cross-correlogram and waveform metrics satisfied criteria of belonging to a single neuron (yet with spikes incorrectly split during spike sorting) were merged. Second, individual units whose auto-correlograms corresponded to more than one neuron (i.e., multi-unit activity) were excluded. After spike sorting and curation, units with an average firing rate below 0.5 Hz within a given day were excluded. All remaining units were included in downstream analyses.

Behavioral Beta-Bernoulli model. To estimate the rats' expected belief about the value of each of the six ports, we used a Beta-Bernoulli model, which estimates the likelihood of any possible value by tracking past successes (rewards) and failures (omissions) at that port. This infers not only the expected value of a given port (analogous to traditional Q learning), but also the probability of the port having *any* possible underlying success rate. In fact, this model provides an optimal estimate of the underlying success rate given the number of past successes and failures, assuming that success rate is constant in time.

The Beta-Bernoulli model tracks port i with two variables, α_i and β_i, which correspond to (with some adjustment, detailed below) the respective number of successes and failures at port i. Alternatively, the model can be re-parameterized via a mean and standard deviation μ_i and σ_i, which eliminates the need to maintain a count of past events, and permits a simple neural implementation.

Port Q value estimation. To estimate the Q value of any given port, expected value of port i is given by $Q_{port_i} = \frac{\alpha_i}{\alpha_i+\beta_i}$, providing a simple estimate to compare ports.

Patch Q value estimation. On each trial, the Q value of the current patch (if the rat were to stay) was computed as a weighted average of its two ports, $Q_{patch} = \gamma * Q_{upcoming} + (1-\gamma)Q_{past}$ where Q_{past} denotes the Q value of the port the rat departed from and $Q_{upcoming}$ denotes the Q value of the port the rat was heading toward. For alternative patches, Q_{patch} was defined as the mean of the two port Q values within each patch.

*Learning rule***.** At the start of each session, α_i and β_i of each port were initialized to 1. After each port visit, the model was updated with the following rule:
After a reward at port i:

$$\alpha_i = \alpha_i + 1$$

After an omission at port i:

$$\beta_i \leftarrow \beta_i + \frac{1-\mu_i}{1/d - \mu_i}$$

where μ_i is the expected value of port i given by $\mu_i = \frac{\alpha_i+1}{\alpha_i+\beta_i+2}$, and d is the current estimation of reward depletion, beginning at 1 and multiplied by a *depletion factor* at each re-visit to the same port. Note that the above role departs slightly from a standard Beta-Bernoulli model where β_i would increment by 1 on a non-reward. This modification correctly accounts for the loss of information due to depletion. Briefly, consider the two extreme cases of d. Upon the first visit to a port, $d = 1$, and the above update rule reduces to the standard rule of $\beta_i \leftarrow \beta_i + 1$. At the other extreme, as d approaches 0 with repeated visits, the reward probability approaches 0, and the resulting update of $\beta_i \leftarrow \beta_i$ presents the fact that a negative outcome provides minimal information about the underlying reward rate. The *depletion factor* was estimated for each day from rat choices and reward outcomes on each trial.

Forgetting. To account for the possibility of forgetting when rats did not visit a port for a while, α_i and β_i were decayed towards 1 after each trial:

$$\alpha_i = (\alpha_i - 1) * decay + 1$$
$$\beta_i = (\beta_i - 1) * decay + 1$$

Behavioral choice model. To connect the value learning model to choices and assess our ability to capture animal behavior, we computed the likelihood of patch choice on every trial as the *softmax* of the three integrated patch values, given by:

$$p_{patch_i} = \frac{e^{V_{patch_i}}}{e^{V_{patch_1}} + e^{V_{patch_2}} + e^{V_{patch_3}}}$$

where the integrated value at the current patch was given by:

$$V_{patch_i} = \beta_{stay} Q_{patch_i} + b_{stay}$$

The value of the patch clockwise from the current patch was given by:

$$V_{patch_i} = \beta_{go} Q_{patch_i} + b_{turn}$$

The value of the remaining patch was given by:

$$V_{patch_i} = \beta_{go} Q_{patch_i}$$

where Q_{patch_i} is the patch Q value estimated by the Beta-Bernoulli model as described above.

The inverse temperature parameters, β_{go} and β_{stay}, determined how sensitive choice behavior was to differences in expected reward – high temperatures indicate high sensitivity to reward, while low temperatures indicate insensitivity. To allow for differential sensitivity to reward estimates at the current patch versus alternative patches, we fit two independent softmax temperatures, β_{stay} for the current patch, and β_{go} for the two alternative patches. The term β_{go} reflected a fixed cost to changing patches due to the time and distance to reach a new patch (as opposed to the reward sensitivity fit by β_{stay} that reflected a tendency to stay in the current patch), and was added to the valuation of the current patch.

To incorporate tendencies of the animals to prefer to turn in certain directions when changing patches, we included a *turn bias* b_{turn}, an offset consistently applied to the patch to the left of the animal's current position. If the rat was in patch A, this was added to patch B. If the rat was in patch B, this was added to patch C, and if the rat was in patch C, this was added to patch A.

On trials where the animal switched patches, likelihood of port choice was modeled as a subsequent softmax between the two ports in the new patch, with a separate softmax temperature β_{port} applied to the per-port reward estimates:

$$p_{port_{L/R}} = \frac{e^{V_{port_{L/R}}}}{e^{V_{port_L}} + e^{V_{port_R}}}$$

The value of the left port was given by:

$$V_{port_L} = \beta_{port} Q_{port_L}$$

And the value of the right port was given by:

$$V_{port_R} = \beta_{port} Q_{port_R}$$

Behavioral model fitting procedure. For the Beta-Bernoulli model, we optimized the free parameters ($including\ \gamma, depletion\ factor, decay, \beta_{go}, \beta_{stay}, \beta_{port}, b_{stay}\ and\ b_{turn}$) of the model by embedding each of them within a hierarchical model to allow parameters to vary from day-to-day. These day-level parameters were themselves modeled as arising from a population-level Gaussian distribution over days, separately for each rat. We estimated the model for each rat to obtain optimal day- and rat-level parameters that minimized the negative log likelihood of predicting the actual choice using an expectation maximization (EM) algorithm with a Laplace approximation to the day-level marginal likelihoods in the M step[48].

Definition of switch value. We defined the *switch value* on each trial as the difference between the highest-valued alternative patch and the current patch,

$$\text{Switch value} = \max(V_{alt.\ patch}) - V_{curr.\ patch}$$

where patch values V_{patch_i} were estimated as described above. Larger switch values therefore indicate greater relative incentive to leave the current patch. Note that switch values were

consistently negative because the behavioral model added a patch-stay bias to the value of the current patch to capture the cost of patch switching.

Generalized Linear Models (GLMs) for single-unit coding of task structure and value. We quantified how individual mPFC neurons encoded task structure and switch value using a previously established Poisson GLM framework[32,49]. In this framework, neuronal spikes are modeled as a Poisson process whose time-varying rate is a log-linear function of covariates (task and behavioral variables) during each regression bin. Because trials were self-initiated and varied in duration, we discretized each of the six stay journeys into eight path-progression bins rather than using fixed-duration temporal bins (i.e., 48 journey × progression bins total). For each neuron, we computed spike counts within each progression bin on each trial. We then defined sets of covariates to capture (1) path progression in each journey using a binary indicator that represents the presence or absence of the rat in a specific path progression bin, and (2) the interaction between path progression and switch value:

$$r = e^{\beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{j=1}^{m} \gamma_j v_j}$$

where r denotes a vector of firing rates for one neuron over T time points, x_i is a binary vector indicating whether an animal was in progression bin i (out of n bins total) at time t, and v_j is a vector indicating the behaviorally-inferred switch value when the animal was in progression bin j (out of m bins total). To assess how broadly each neuron's value coding generalized across task conditions (Extended Data Fig. 1B), we used the same basic GLM structure and retained the journey × progression bins (x_i for i = 1, 2,… 48), but adjusted the set of progression-value interaction terms (v_j) to reflect different condition groupings:

- Patch-level GLM: 24 patch ID × progression value terms
- Action-level GLM: 16 action × progression value terms
- Progression-level GLM: 8 path progression value terms

To assess the significance of full models against null (intercept-only) models or no-value models , we used 10-fold cross validation and repeated 100 times per neuron. Additionally, L1 regularization was applied to prevent overfitting.

GLM comparison and neuron categorization (Extended Data Fig. 1B). We determined the level of condition generality at which each neuron encoded switch value by pairwise comparison between GLMs fitted with (i) value × progression terms, (ii) value × progression × action terms, (iii) value × progression × patch ID terms, or (iv) value × progression × journey terms. All models included the same set of progression × journey indicators (i.e., the no-value terms), ensuring that only the value-modulation structure varied across models. For each neuron, we compared the fraction of deviance explained by each pair of GLMs using one-sided Wilcoxon rank-sum test, with Holm–Bonferroni correction applied across comparisons. A neuron was assigned a value-coding generality category based on the following criteria:

- If one model explained significantly more fraction of deviance than all others, that model was selected.
- If multiple models each explained significantly more deviance than the remaining models but showed statistically indistinguishable performance, we selected the most parsimonious one among them.

This procedure yielded a principled categorization of each neuron according to the best-fitting and least complex model of its value-encoding generality.

Principal component analysis (PCA) and subspace definition. To explore population-level structure in mPFC activity of each day, we used PCA to identify neural dimensions capturing the largest variance within a specified analysis window (Extended Data Fig. 2). Specifically, neuronal firing rates were estimated by convolving spike trains of each single unit with a Gaussian kernel (σ = 50 ms). We then constructed a data matrix R (Conditions × Time) × Neurons, where conditions were defined by proximity to patch switches (Fig. 3A, C), proximity to patch switches × action (Fig. 3B, Extended Data Fig. 3C), proximity to patch switches × patch identity (Extended Data Fig. 3A, D), or proximity to patch switches × journey (Extended Data Fig. 3B, E). Firing rates of each neuron were averaged across all trials belonging to a given condition. The **navigation subspace** was identified by applying PCA to neural data from 0 to +3 s relative to trial initiation (i.e., the last nose-poke time at the starting port). This interval covered the median run-time period on stay trials. We then projected neural data from the same time window into this subspace (Fig. 3A, B; Extended Data Fig. 3A, B). To isolate dimensions more strongly linked to value-related variance, we performed a second PCA on neural activity from –0.8 to +0.2 s relative to trial initiation, an interval during which rats paused at reward ports on both rewarded and unrewarded trials before initiating the next movement (Extended Data Fig. 4). The leading PCs from this decomposition defined the **pre-move subspace**. We then projected neural data (i) during the -0.8 to +3 s window relative to trial initiation on stay trials and the -0.8 to +4.5 s window relative to trial initiation on switch trials into this subspace to reveal the continuous spiral trajectories across trials (Fig. 3C, Extended Data Fig. 3C-E), and (ii) during the -5 to 0 s interval from trial initiation after reward and the -0.8 to 0 s interval from trial initiation after no reward to reveal evolution of neural trajectories spanning the outcome period of the preceding trial (Fig. 3D).

Cross-validated switch value decoding from neural population activity. To quantify value representations in mPFC population activity, we decoded the *switch value* estimated by the Beta–Bernoulli behavioral model using LASSO regression with L1 regularization. Given the interaction between task structure and value in neuronal coding, we discretized each of the six stay journeys into multiple progression bins per trial and averaged neuronal spiking data within each bin. Depending on the level of condition generality examined, bins were defined as:

- Progression-based value decoder (Fig. 4; Extended Data Fig. 6): all stay trials were pooled together and divided into 11 progression bins each (8 bins during run time and 3 bins during pause time).
- Action-based value decoder (Extended Data Fig. 6D): stay trials with the same actions (turning left or turning right) were grouped together and trials in each group were divided into 8 run-time progression bins (i.e., 16 progression × action bins total).
- Patch-based value decoder (Extended Data Fig. 6D): stay trials within the same patch were grouped together and trials in each group were divided into 8 run-time progression bins (i.e., 24 progression × patch bins total).
- Journey-based value decoder (Extended Data Fig. 6D): stay trials of the same journey type were grouped together and trials in each group were divided into 8 run-time progression bins (i.e., 48 progression × journey bins).

For each bin, we trained a decoder of the form:

$$V = \beta_0 + \sum_{i=1}^{N} \beta_i r_i$$

where r_i denotes the firing-rate vector of neuron i (out of N simultaneously recorded neurons) across all time points assigned to that bin, β_i is its decoding weight, and V is the corresponding vector of Beta-Bernoulli value estimates. All bins belonging to the same trial were assigned the same per-trial value estimate, so repeated entries in V reflect multiple time points from a single trial rather than moment-by-moment value updates.

To prevent overfitting, we implemented a leave-one-out cross-validation procedure: we grouped all trials by their proximity to a patch-switch decision (e.g., six visits away, five visits away, so on and so forth, from an upcoming switch) and used all trials except those in one of these groups to train the model; the held-out group (e.g., all trials that were two visits away from a switch) was used exclusively for testing. Critically, this procedure was repeated so that each stay-trial group had been left out once and all switch trials were always excluded from training. To compare decoding performance across decoders at different levels of condition generality (Extended Data Fig. 6D), we randomly subsampled trials when training the progression-based, action-based, and patch-based value decoders so that all models were fitted on matched numbers of trials. This ensured that differences in decoding accuracy reflected structure in the neural data rather than disparities in sample size.

Quantification of neural value resetting. To measure the degree to which neural value representations 'reset' when rats entered a new patch (Fig. 5A-C), we computed:

Relative Neural Resetting = First-stay value – Switch-in value

where *Switch-in value* is the average decoded switch value across the four post-choice bins when rats switched to a new patch, and *First-stay value* refers to the average decoded switch value across the four post-choice bins of the subsequent first stay trials in that same patch. We restricted analyses to cases where the switch-in trials were rewarded to ensure comparability across early and late learning.

Quantification of neural value updating. To assess how reward outcomes influenced value representation, we defined two complementary metrics: (i) *'Outcome effect'* on switch value was quantified as *Outcome value – Final switch-in value*, where *Final switch-in value* is the decoded switch value in the final run-time bin (bin 8) of a switch trial, and *Outcome value* denotes the mean decoded value in the outcome bins (pause period) of that same switch trial (Fig. 5D); (ii) *'Neural value update'* was quantified as *Initial first-stay value – Final switch-in value*, where *Final switch-in value* is as defined in (i), and *Initial first-stay value* denotes the decoded value in bin 1 of the subsequent stay trial (Fig. 5D). We then assessed the effect of reward outcomes on *Outcome effect* and *Neural value update* by computing:

(i) ΔOutcome effect = mean(Outcome effect|Reward=0) − mean(Outcome effect|Reward=1)
(ii) ΔValue update = mean(Value update|Reward=0)−mean(Value update|Reward=1)

which are shown in Fig. 5E.

Predicting choices from behavioral value estimates or neural value decoding. To quantify how well value variables predicted rats' decisions, we fit logistic regression models (GLMs with a logit link) to classify stay versus switch choices using either the switch value estimated by the behavioral Beta-Bernoulli model or the switch value decoded from neural population activity (Figs. 1M, 4D; Extended Data Fig. 6C). Models were fit separately for each animal. Each model used a single

predictor—the behaviorally estimated or neurally decoded switch value of each trial—and output the probability of a switch decision. Model predictions were binarized using a decision threshold optimized to maximize overall prediction accuracy. To avoid biased evaluation of decoding accuracy due to unbalanced numbers of switch and stay trials, switch-decoding accuracy (sensitivity) and stay-decoding accuracy (specificity) were computed separately and then averaged (Fig. 4D, Extended Data Fig. 6C).

Statistics

Statistical tests for each analysis are described in the corresponding Methods subsections and figure legends. Unless otherwise noted, all bootstrapping was done over 10,000 iterations. All model comparisons were performed on models trained with matched numbers of trials, achieved by random subsampling to ensure equivalent sample sizes. Statistical significance was assessed at $\alpha = 0.05$, with significance levels indicated as: * $P < 0.05$, ** $P < 0.01$, and *** $P < 0.001$.

Data and code availability

Data and code will be shared publicly via the Distributed Archives for Neurophysiology Data Integration (DANDI) Archive and a repository at https://github.com/LorenFrankLab. Further information necessary for analysis of this dataset is available upon request from the corresponding author(s).

Acknowledgements

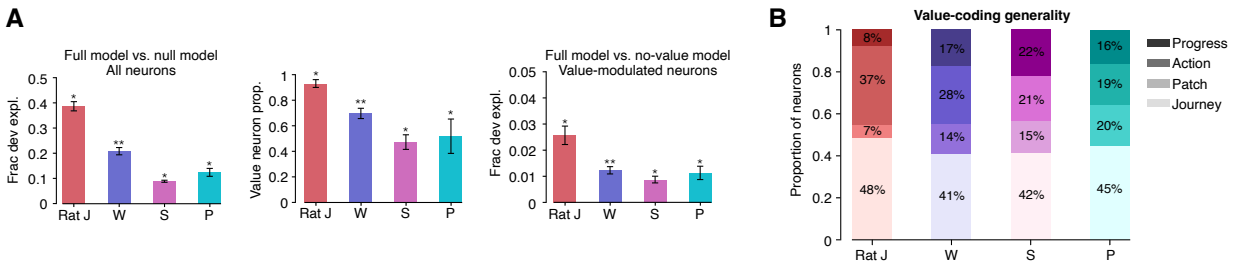
We thank A. Kiseleva, V. Perez, and M. Yusifova for administrative support, V. Kharazia for assistance with histology, L. Tong for assistance with spike curation, S. Gu for assistance with behavioral modeling, C. B. Washington and C. Pandarinath for assistance with single-trial analysis, S. Tseng, D. J. O'Shea, S. Vyas, and L. Tian for analysis discussions, and A. K. Gillespie for rat illustrations. This work was supported by the Jane Coffin Childs Memorial Fund (X.S.), NSF GRFP 1650113 (A.E.C.), NIH F31MH124366 (A.E.C.), UCSF Jonas Cohler Discovery Fellowship (A.E.C.), the Life Sciences Research Foundation (A.J.), NIH F30MH126483 (J.A.G.), NIH R01MH121093 (N.D.D.), NIH R01MH136875 (J.D.B., N.D.D., L.M.F.), Simons Collaboration on the Global Brain 521921 and 542981 (L.M.F.), and Howard Hughes Medical Institute (L.M.F.).

Author contributions

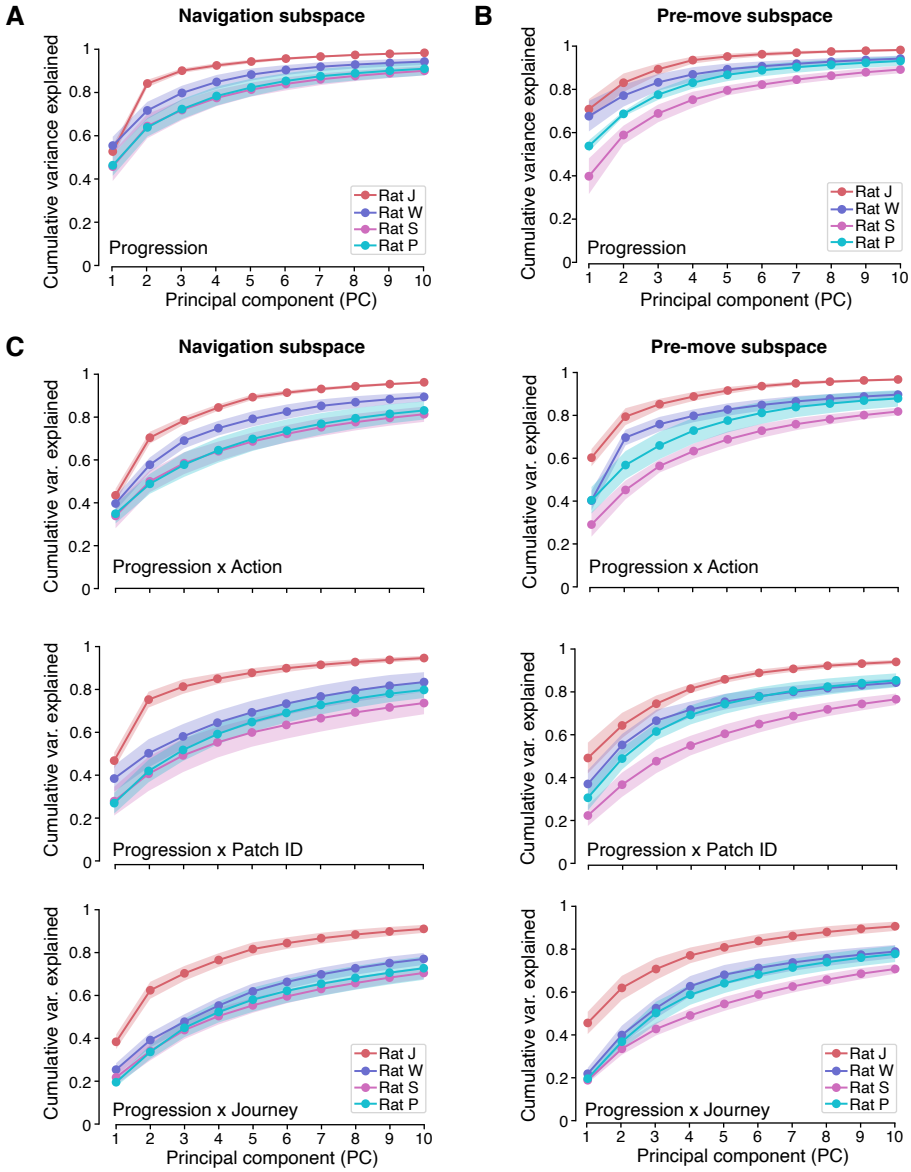
Conceptualization, X.S., A.E.C., N.D.D., L.M.F.; Methodology, X.S., A.E.C., A.E.K., T.A.K., J.D.B., N.D.D., L.M.F.; Data Collection, A.E.C., E.J.M.; Surgical Procedures, A.E.C., E.J.M., A.J., J.A.G.; Formal Analysis, X.S., A.E.C., A.E.K.; Software, X.S., A.E.C., A.E.K., E.J.M., E.L.D., N.D.D., L.M.F.; Resources, J.Z., P.T., J.H., A.Y., R.H., L.M.F.; Data Curation, X.S., A.E.C., E.J.M., L.M.F.; Writing – Original Draft, X.S., A.E.K., L.M.F.; Writing – Review & Editing, X.S., A.E.C., A.E.K., E.J.M., A.J., J.A.G., T.A.K., E.L.D., J.D.B., N.D.D., L.M.F.; Visualization, X.S., A.E.K., A.E.C.; Supervision, L.M.F.

Declaration of interests

The authors declare no competing interests.

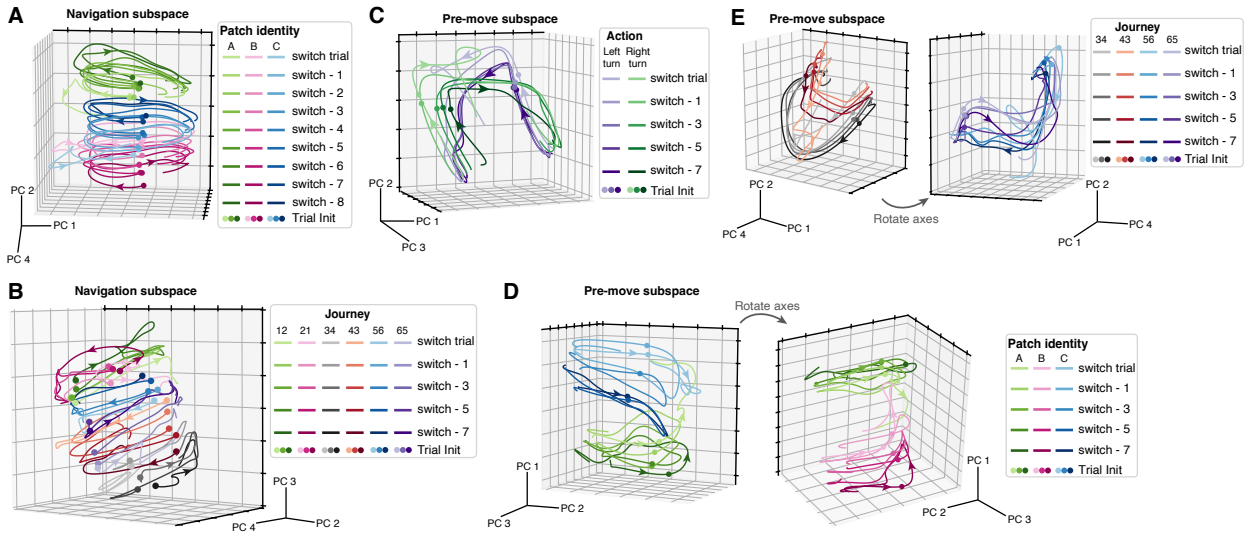


Extended Data Fig. 1 | Quantifications of GLM coefficients and significance.
A, GLM summary statistics across all days for each rat. Left: The full GLM (including value terms) explained significantly more fraction of deviance in single-neuron activity than the null (intercept-only) model (one-sided Wilcoxon signed-rank test > 0: $P_J = 0.016$, $P_W = 0.0078$, $P_S = 0.016$, $P_P = 0.031$). Middle: Proportion of value-encoding neurons (one-sided Wilcoxon signed-rank test > 0: $P_J = 0.016$, $P_W = 0.0078$, $P_S = 0.016$, $P_P = 0.031$). Right: Fraction of deviance attributable specifically to value terms in the full model was significantly greater than zero (one-sided Wilcoxon signed-rank test > 0: $P_J = 0.016$, $P_W = 0.0078$, $P_S = 0.016$, $P_P = 0.031$). Error bars: standard deviation. **B,** Categorization of value-modulated neurons by their value-coding generality. For each neuron, we performed pairwise comparisons among GLMs incorporating value × progression, value × progression × action, value × progression × patch ID, or value × progression × journey terms. The most parsimonious model whose fraction of deviance explained was not significantly lower than any competing model defined the neuron's value-coding generality (one-sided Wilcoxon rank-sum test with Holm-Bonferroni correction; see Methods).

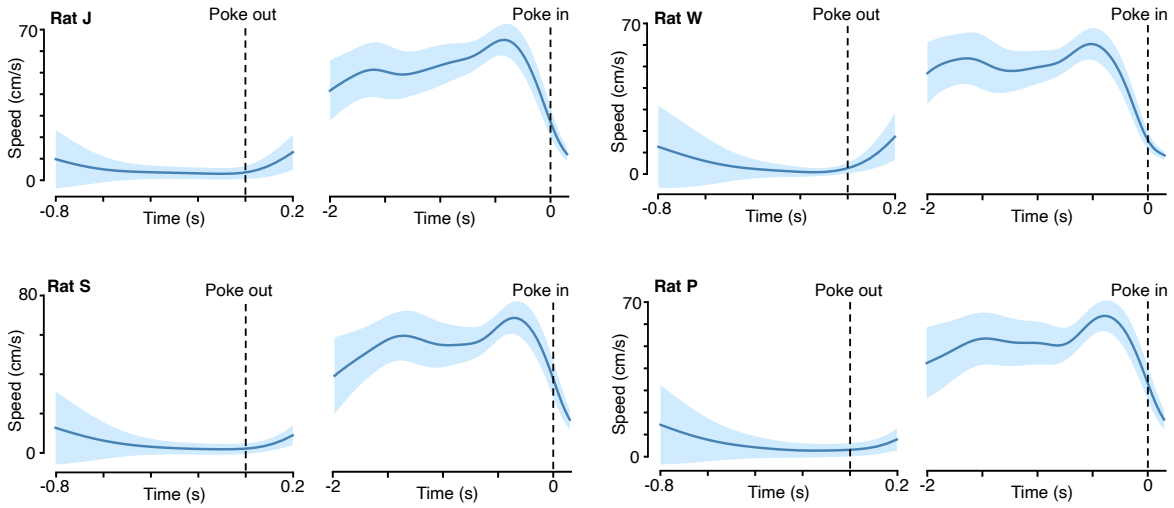


Extended Data Fig. 2 | Variance explained by navigation and pre-move neural subspaces constructed by PCA.

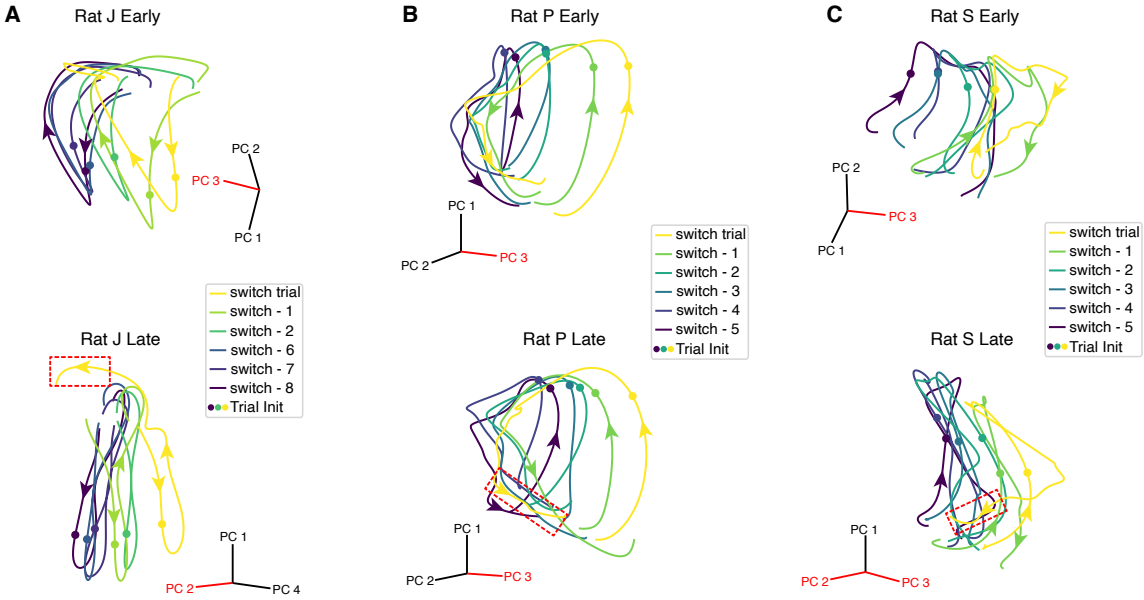
A, Cumulative variance explained by the leading PCs of run-time neural data grouped by proximity to a switch trial. **B,** Cumulative variance explained by the leading PCs of pre-move (a subset of the pause period) neural data grouped by proximity to a switch trial. **C,** Cumulative variance explained by the leading PCs of run-time and pre-move neural data grouped by different conditions, including action × switch proximity (top), patch ID × switch proximity (middle), and journey × switch proximity (bottom). **A-C,** Shaded areas: standard deviation across all days for each rat.



Extended Data Fig. 3 | Generalized dynamical motifs across task conditions.
A, In the navigation subspace, progression-tracking trajectories for each foraging patch were orderly displaced according to proximity to switch, correlated with the switch value gradient. **B,** Progression-tracking trajectories similarly sorted by switch proximity for each individual journey in the navigation subspace. Every other pre-switch trial is shown for clearer visualization. **C,** In the pre-move subspace, neural trajectories expressed a value × progression × action motif, with value-related orderly displacement across pre-switch trials. Every other pre-switch trial is shown for clearer visualization. **D,** Dynamical motifs tracking value × progression × patch ID in the pre-move subspace. Showing every other pre-switch trial and two patches at a time for clearer visualization. **E,** Dynamical motifs tracking value × progression × journey in the pre-move subspace. Showing every other pre-switch trial and two pairs of journeys for clearer visualization. **A-E,** All sessions of one example day are shown. Circles: trial initiation. Arrowheads: direction of neural trajectory evolution.

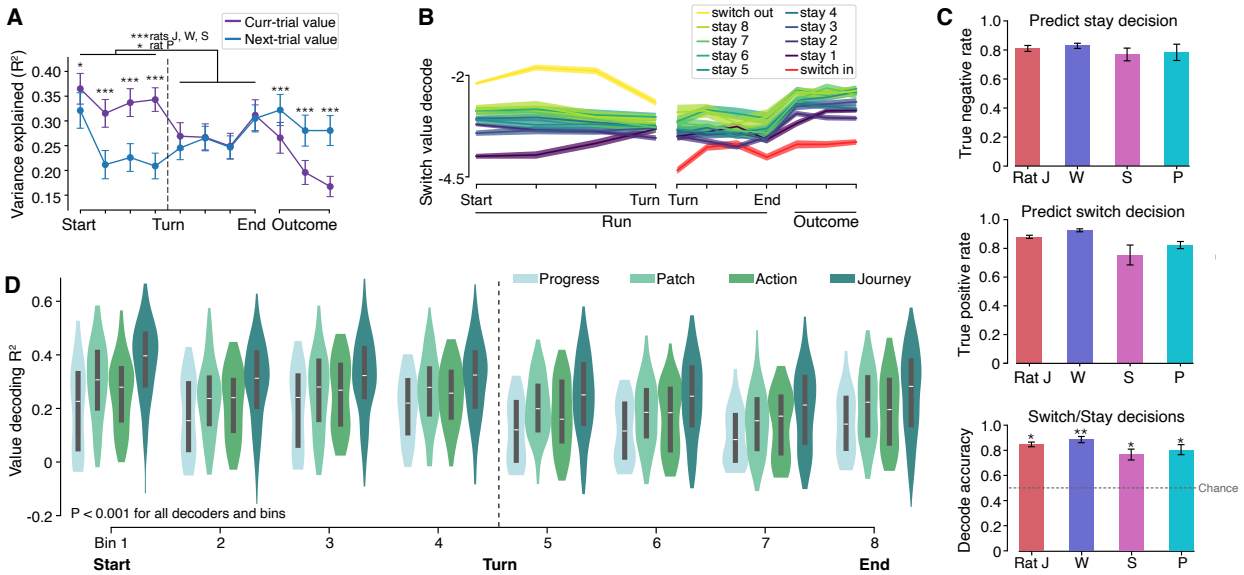


Extended Data Fig. 4 | Rats' movement speed during pre-move and run periods.
Left panel of each rat: Mean speed remained at or below ~10 cm/s throughout the pre-move time window used to construct the pre-move neural subspace, indicating minimal locomotion. Such immobility during this period allowed extraction of neural variance primarily related to value. 'Poke out' marks the time when rats withdrew from the reward port. Right panel of each rat: During the run period of each trial, mean speed remained above ~30 cm/s until rats approached the reward port, consistent with sustained locomotion throughout this interval. 'Poke in' marks the moment rats arrived at the port. Shaded areas: standard deviation across all days for each rat.

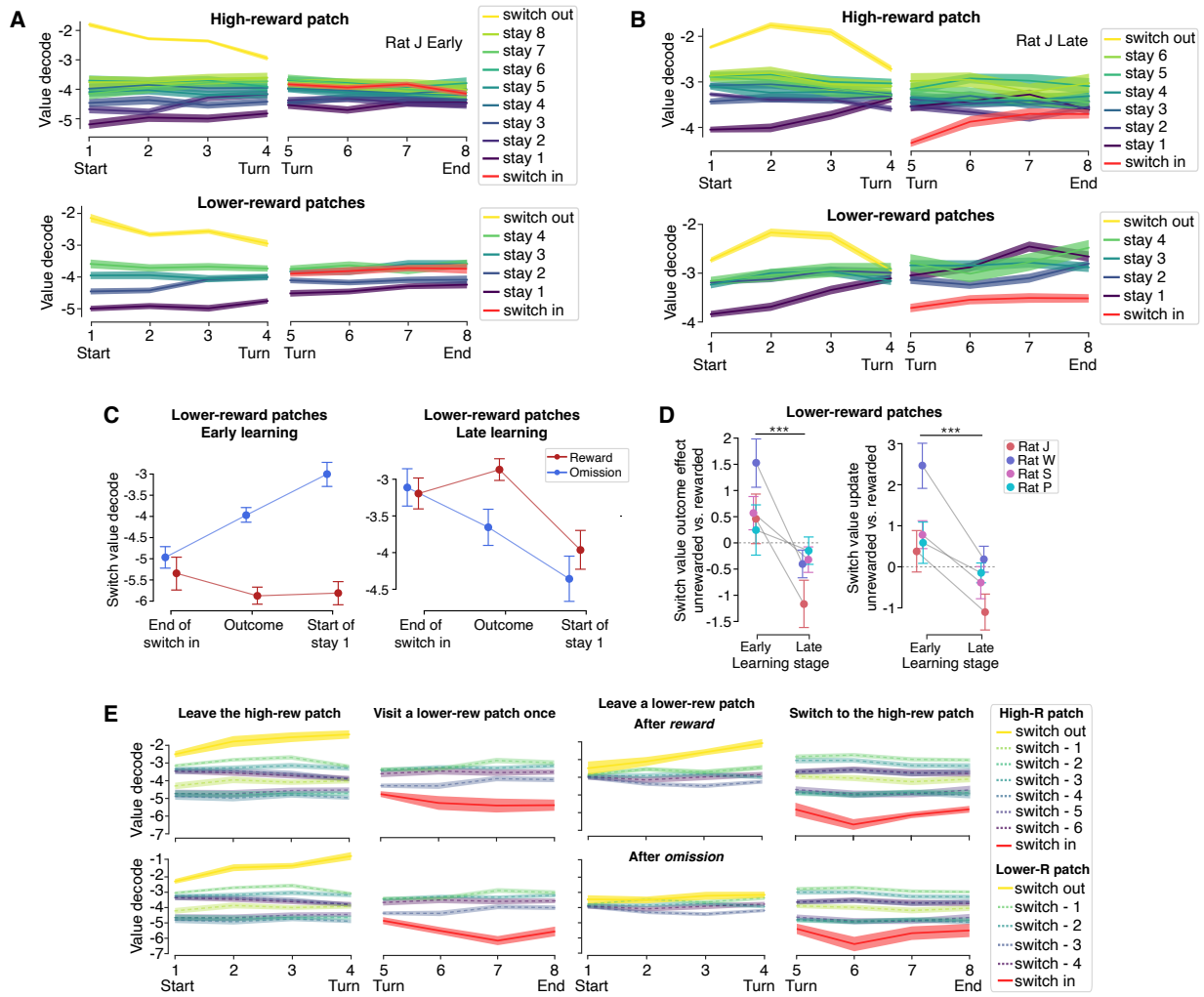


Extended Data Fig. 5 | Trial-averaged neural trajectories in the pre-move subspace during early and late learning.

A-C, Trial-averaged neural trajectories grouped by proximity to a switch decision. For each rat, two example days are shown. The PC axes along which we observed prominent value-related displacement of spiral trajectories are highlighted in red. The resetting trajectories corresponding to entry into the final arm of a new patch (before reward outcome delivery) are emphasized inside rectangular boxes. In early-learning sessions, rats J and S showed minimal resetting upon entering a new patch, whereas rats W (see Fig. 3C) and P exhibited partial resetting that was weaker than in late-learning sessions. Note that although PCA is useful for visualizing dynamical structure, variability in trial counts and the number of recorded neurons can affect the apparent geometry of low-dimensional trajectories (e.g., occasional entangling of spiral gradients in late sessions for rats J and S). Importantly, our conclusions primarily relied on decoding-based representational analyses, which provided a more robust quantification of value-coding structure. Nevertheless, PC-space trajectories remain informative for interpreting the organization and evolution of underlying neural dynamics.

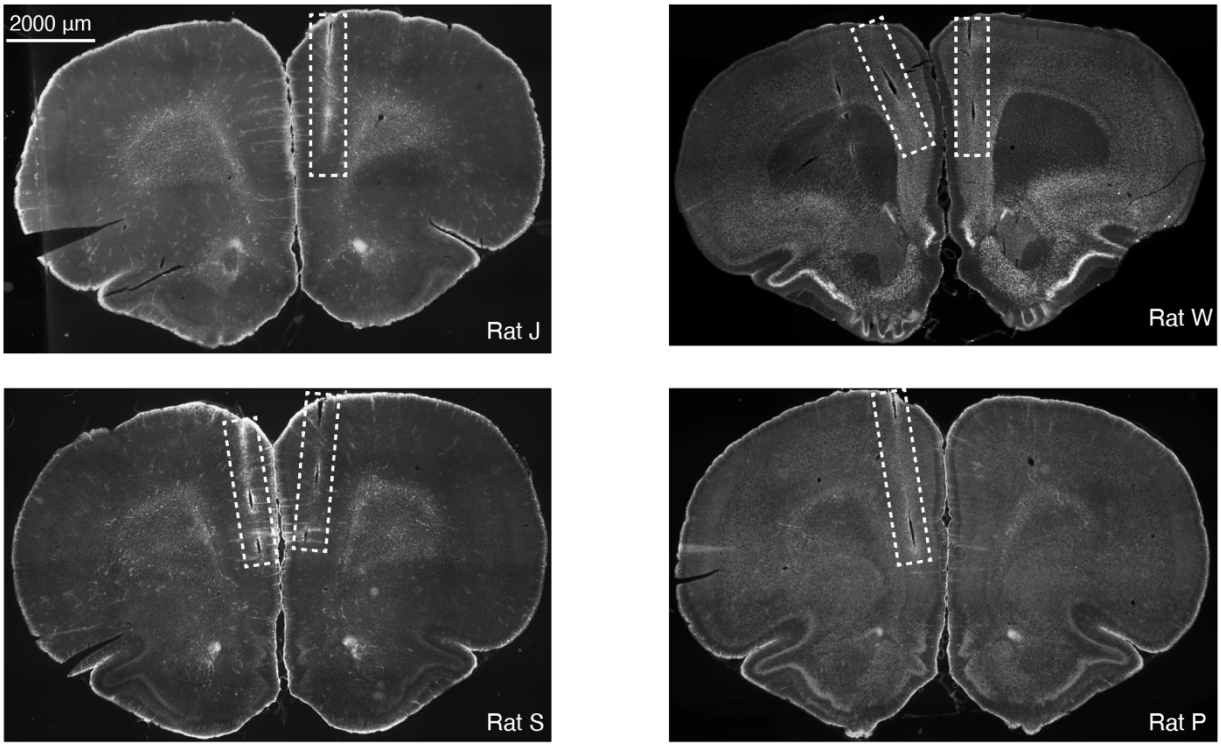


Extended Data Fig. 6 | Additional measurements of switch value decoding.
A, Variance in the switch value (estimated by the behavioral model) explained by the neural decoder across 11 progression bins. Data from all sessions and rats were pooled for statistical testing. In pre-choice bins (1–4), the decoder captured significantly more variance in the *current-trial value* than in the *next-trial value* (paired one-sided Wilcoxon rank-sum test: $P_{bin1} = 0.011$, $P_{bin2} = 6.35 \times 10^{-5}$, $P_{bin3} = 3.28 \times 10^{-6}$, $P_{bin4} = 5.96 \times 10^{-8}$). The relationship reversed in outcome-period bins (9–11) , with the decoder explaining more variance in the next-trial value (paired one-sided Wilcoxon rank-sum test: $P_{bin9} = 4.17 \times 10^{-6}$, $P_{bin10} = 1.19 \times 10^{-7}$, $P_{bin11} = 5.96 \times 10^{-8}$). When decoding current-trial value, variance explained was higher in pre-choice than post-choice bins (permutation test per animal: $P_J = 0.0$, $P_W = 0.0$, $P_S = 0.0006$, $P_P = 0.0248$). **B,** Decoded switch value across 11 progression bins for stay and switch trials in an example day (5 sessions). Value during the outcome period appeared to link the terminal value of one trial to the initial value of the next, consistent with value updating. Shaded areas: SEM across trials. **C,** Decoding accuracy for stay decisions (top), switch decisions (middle), and the mean of both (bottom) across all days for each rat. Accuracy remained stable across days. One-sided Wilcoxon signed-rank test > 0.5 (bottom): $P_J = 0.016$, $P_W = 0.0078$, $P_S = 0.016$, $P_P = 0.031$. Error bars: standard deviation. **D,** Variance in the switch value (estimated by the behavioral model) explained by neural decoders trained at different levels of condition generality (during run time): 8 progression bins, 16 action × progression bins, 24 patch × progression bins, and 48 journey × progression bins. We subsampled trials before fitting the progression-, action-, and patch-based decoders to ensure that the performance of all decoders were comparable with matched training and test data sizes. Variance in the switch value explained by all decoders was significantly greater than zero (all days of all rats pooled for statistical testing; one-sided Wilcoxon signed-rank test: $P < 0.001$ across all progression bins, action × progression bins, patch × progression bins, and journey × progression bins).



Extended Data Fig. 7 | Changes in neural value representations at finer timescales and in lower-reward patches are consistent with the meta-learned strategy.

A, B, Decoded switch value across run-time progression bins within a patch for switch-in (red), the subsequent stay (violet to green), and switch-out (yellow) trials. Early- and late-learning sessions of an example rat are shown (same datasets as in Fig. 5A, B). Trials in the high-reward and lower-reward patches were analyzed separately because their average bout lengths typically differed (Fig. 1I). Decoded switch value during switch-in trials reset substantially more in late-learning sessions than in early-learning sessions. Shaded areas: SEM across trials. **C, D,** Same measurements as in Fig. 5D, E for lower-reward patches. **D,** Linear mixed effects model testing random slopes per animal: outcome-effect difference—$P_J = 9.05 \times 10^{-6}$, $P_W = 4.97 \times 10^{-12}$, $P_S = 6.78 \times 10^{-4}$, $P_P = 0.18$; value-update difference—$P_J = 4.35 \times 10^{-5}$, $P_W = 5.78 \times 10^{-12}$, $P_S = 3.19 \times 10^{-4}$, $P_P = 0.0075$. Fixed slope across animals: outcome-effect difference, $P = 3.80 \times 10^{-14}$; value-update difference, $P = 3.13 \times 10^{-14}$. Error bars: 95% CI. **E,** Decoded switch value in run-time progression bins during single-visit episodes (solid curves). For comparison, decoded switch value during stay bouts (i.e., when rats stayed for more than a single visit) in the same patches are shown as washed out, dashed curves. After both rewarded and unrewarded single visits, decoded switch value at the start of the subsequent trial was, on average, higher than that during stay bouts (third column). Shaded areas: SEM across trials.



Extended Data Fig. 8| Implant locations. Example PFC histological sections from rats J, W, S, and P. Tracts of the mPFC probes are emphasized inside rectangular boxes.

References

1. Mobbs, D., Trimmer, P. C., Blumstein, D. T. & Dayan, P. Foraging for foundations in decision neuroscience: insights from ethology. *Nat. Rev. Neurosci.* **19**, 419–427 (2018).
2. Kolling, N., Behrens, T. E. J., Mars, R. B. & Rushworth, M. F. S. Neural mechanisms of foraging. *Science* **336**, 95–98 (2012).
3. Harlow, H. F. The formation of learning sets. *Psychol. Rev.* **56**, 51–65 (1949).
4. Harlow, H. F. & Warren, J. M. Formation and transfer of discrimination learning sets. *J. Comp. Physiol. Psychol.* **45**, 482–489 (1952).
5. Griffiths, T. L. *et al.* Doing more with less: meta-reasoning and meta-learning in humans and machines. *Curr. Opin. Behav. Sci.* **29**, 24–30 (2019).
6. Zhao X., Gattoni R., Kozlosky A., Jacobs A., Morrow C, Lindo S. & Spruston N. Meta-learning in head fixed mice navigating in virtual reality: A Behavioral Analysis. *bioRxiv* (2023) doi:10.1101/2023.05.01.538936.
7. Nam, A. J. & McClelland, J. L. Systematic human learning and generalization from a brief tutorial with explanatory feedback. *Open Mind* **8**, 148–176 (2024).
8. Achterberg, J. *et al.* A one-shot shift from explore to exploit in monkey prefrontal cortex. *J. Neurosci.* **42**, 276–287 (2022).
9. Wang, J. X. *et al.* Prefrontal cortex as a meta-reinforcement learning system. *Nat. Neurosci.* **21**, 860–868 (2018).
10. Hattori, R. *et al.* Meta-reinforcement learning via orbitofrontal cortex. *Nat. Neurosci.* **26**, 2182–2191 (2023).
11. Wang, J. X. Meta-learning in natural and artificial intelligence. *Curr. Opin. Behav. Sci.* **38**, 90–95 (2021).
12. Goudar, V., Peysakhovich, B., Freedman, D. J., Buffalo, E. A. & Wang, X.-J. Schema formation in a neural population subspace underlies learning-to-learn in flexible sensorimotor problem-solving. *Nat. Neurosci.* **26**, 879–890 (2023).
13. Samborska, V., Butler, J. L., Walton, M. E., Behrens, T. E. J. & Akam, T. Complementary task representations in hippocampus and prefrontal cortex for generalizing the structure of problems. *Nat. Neurosci.* **25**, 1314–1326 (2022).
14. Hospedales, T., Antoniou, A., Micaelli, P. & Storkey, A. Meta-learning in neural networks: A survey. *arXiv [cs.LG]* (2020).
15. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
16. Jensen, K. T., Hennequin, G. & Mattar, M. G. A recurrent network model of planning explains hippocampal replay and human behavior. *Nat. Neurosci.* **27**, 1340–1348 (2024).
17. Bromberg-Martin, E. S., Matsumoto, M., Hong, S. & Hikosaka, O. A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.* **104**, 1068–1076 (2010).
18. El-Gaby, M. *et al.* A cellular basis for mapping behavioural structure. *Nature* **636**, 671–680 (2024).
19. Bernardi, S. *et al.* The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell* **183**, 954-967.e21 (2020).
20. Genovesio, A., Brasted, P. J., Mitz, A. R. & Wise, S. P. Prefrontal cortex activity related to abstract response strategies. *Neuron* **47**, 307–320 (2005).
21. Lake, B. M. & Baroni, M. Human-like systematic generalization through a meta-learning neural network. *Nature* **623**, 115–121 (2023).
22. Guidera, J. A. *et al.* Regional specialization manifests in the reliability of neural population codes. *bioRxivorg* (2024) doi:10.1101/2024.01.25.576941.

23. Bari, B. A. *et al.* Stable representations of decision variables for flexible behavior. *Neuron* **103**, 922-933.e7 (2019).
24. Akam, T. *et al.* The anterior cingulate cortex predicts future states to mediate model-based action selection. *Neuron* **109**, 149-163.e7 (2021).
25. Klein-Flügge, M. C., Bongioanni, A. & Rushworth, M. F. S. Medial and orbital frontal cortex in decision-making and flexible behavior. *Neuron* **110**, 2743–2770 (2022).
26. Comrie, A. E. *et al.* Hippocampal representations of alternative possibilities are flexibly generated to meet cognitive demands. *bioRxivorg* (2024) doi:10.1101/2024.09.23.613567.
27. Chung, J. E. *et al.* High-density, long-lasting, and multi-region electrophysiological recordings using polymer electrode arrays. *Neuron* **101**, 21-31.e5 (2019).
28. Driscoll, L. N., Shenoy, K. & Sussillo, D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nat. Neurosci.* **27**, 1349–1363 (2024).
29. Yang, G. R., Joglekar, M. R., Song, H. F., Newsome, W. T. & Wang, X.-J. Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* **22**, 297–306 (2019).
30. Frankland, S. M. & Greene, J. D. Concepts and compositionality: In search of the brain's language of thought. *Annu. Rev. Psychol.* **71**, 273–303 (2020).
31. Hattori, R., Danskin, B., Babic, Z., Mlynaryk, N. & Komiyama, T. Area-specificity and plasticity of history-dependent value coding during learning. *Cell* **177**, 1858-1872.e15 (2019).
32. Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P. & Brown, E. N. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* **93**, 1074–1089 (2005).
33. Doll, B. B., Simon, D. A. & Daw, N. D. The ubiquity of model-based reinforcement learning. *Curr. Opin. Neurobiol.* **22**, 1075–1081 (2012).
34. Badre, D., Kayser, A. S. & D'Esposito, M. Frontal cortex and the discovery of abstract action rules. *Neuron* **66**, 315–326 (2010).
35. Strange, B. A., Henson, R. N., Friston, K. J. & Dolan, R. J. Anterior prefrontal cortex mediates rule learning in humans. *Cereb. Cortex* **11**, 1040–1046 (2001).
36. Wallis, J. D., Anderson, K. C. & Miller, E. K. Single neurons in prefrontal cortex encode abstract rules. *Nature* **411**, 953–956 (2001).
37. Reinert, S., Hübener, M., Bonhoeffer, T. & Goltstein, P. M. Mouse prefrontal cortex represents learned rules for categorization. *Nature* **593**, 411–417 (2021).
38. Hattori, R. & Komiyama, T. Context-dependent persistency as a coding mechanism for robust and widely distributed value coding. *Neuron* **110**, 502-515.e11 (2022).
39. Miranda, B. *et al.* Neural signatures of model-based and model-free reinforcement learning across prefrontal cortex and striatum. *bioRxiv* (2025) doi:10.1101/2025.01.11.632388.
40. Miller, K. J., Botvinick, M. M. & Brody, C. D. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* **20**, 1269–1276 (2017).
41. Knudsen, E. B. & Wallis, J. D. Closed-loop theta stimulation in the orbitofrontal cortex prevents reward-based learning. *Neuron* **106**, 537-547.e4 (2020).
42. Schweighofer, N. & Doya, K. Meta-learning in reinforcement learning. *Neural Netw.* **16**, 5–9 (2003).
43. Grima, L. L., Guo, Y., Narayan, L., Hermundstad, A. M. & Dudman, J. T. A global dopaminergic learning rate enables adaptive foraging across many options. *bioRxiv* (2024) doi:10.1101/2024.11.04.621923.
44. Krausz, T. A. *et al.* Dual credit assignment processes underlie dopamine signals in a complex spatial environment. *Neuron* **111**, 3465-3478.e7 (2023).
45. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15**, 495–506 (2002).

46. Lee, K. H. *et al.* Spyglass: a framework for reproducible and shareable neuroscience research. *bioRxiv* (2024) doi:10.1101/2024.01.25.577295.
47. Chung, J. E. *et al.* A fully automated approach to spike sorting. *Neuron* **95**, 1381-1394.e6 (2017).
48. Huys, Q. J. M. *et al.* Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput. Biol.* **7**, e1002028 (2011).
49. McCullagh, P. & Nelder, J. A. An outline of generalized linear models. in *Generalized Linear Models* 21–47 (Routledge, 2019). doi:10.1201/9780203753736-2.