Reinforcement-based option competition in human dorsal stream during exploration/exploitation of a continuous space

Michael N. Hallquist^{†1}, Kai Hwang², Beatriz Luna³, Alexandre Y. Dombrovski*³

¹ Department of Psychology, University of North Carolina, Chapel Hill, NC, USA ² Department of Psychological and Brain Sciences, Iowa Neuroscience Institute, University of Iowa, Iowa City, IA, USA

³ Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA

⁺ A.Y.D. and M.N.H. contributed equally.

<u>* Corresponding author and lead contact</u>: Alexandre Y. Dombrovski, Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, 15213. Email: <u>dombrovskia@gmail.com</u>

Figures: 6

Word count including figure legends excluding Methods: 6452 words

Abstract

Primates exploring and exploiting a continuous sensorimotor space rely on maps in the dorsal stream that guide visual search, locomotion, and grasp. For example, an animal swinging from one tree limb to the next uses rapidly evolving sensorimotor representations to decide when to harvest a reward. We show that such exploration/exploitation depends on dynamic maps of competing option values in the human dorsal stream. Using a reinforcement learning (RL) model capable of rapid learning and efficient exploration and exploitation, we show that preferred options are selectively maintained on the map while the values of spatiotemporally distant alternatives are compressed. Consistent with biophysical models of cortical option competition, dorsal stream BOLD signal increased and posterior cortical β_1/α oscillations desynchronized as the number of potentially valuable options grew, matching predictions of information-compressing RL rather than traditional RL that caches long-term values. BOLD and β_1/α responses were correlated and predicted the successful transition from exploration. These option competition dynamics were observed across parietal and frontal dorsal stream regions, but not in the occipito-temporal MT+ sensitive to the average reward rate. Our results also illustrate that models' diverging predictions about information dynamics can help to adjudicate between them based on population activity.

Graphical abstract



Introduction

Organisms face a difficult dilemma between exploiting options that are known to be good and exploring new options that might be even better. When a vertebrate faces a few discrete options, the striatum and amygdala can resolve the explore-exploit dilemma by representing options egocentrically (e.g., right/left) and tuning the exploration rate based on meso-striatal dopaminergic signals¹⁻⁴. In more complex terrestrial environments, however, quadrupeds rely on world-centric hippocampal cognitive maps that incorporate reinforcement and are stored in long-term memory⁵⁻⁷. While these mechanisms work efficiently at slower timescales, exploration and exploitation become more demanding when we move rapidly through a dynamic environment.

For example, as an adaptation to arboreal hunting and foraging on terminal branches, primates evolved visuo-motor systems that support fast and precise visually guided actions⁸. These behaviors rely on the cortical "where" stream, or the dorsal attention network (DAN), which integrates visual and somatosensory information to build dynamic world-centric maps that guide visual search, locomotion, and grasp. More specifically, the posterior parietal cortex (PPC) constructs maps using visual inputs from temporal-occipital areas such as MT+ as well as parietal somatosensory inputs. In turn, the PPC sends map-based outputs to the frontal dorsal and ventral premotor (PMd and PMv) cortex and the frontal eye fields (FEF) that guide action^{9–11}.

Visuomotor learning has to occur at a faster timescale than instrumental learning in the basal forebrain and striatum, which integrates reinforcement slowly and retains long-term values¹². Moreover, PPC maps contain rich visuomotor data necessary to decide what actions are likely to succeed in current and upcoming spatio-temporal locations, e.g., when an insect can be grasped on a moving branch^{13–15}. These pragmatic maps represent programs of movement toward currently available options that are based on prior visuomotor experience⁹. Studies of gaze control, for example, find that PPC facilitates goal-congruent saccades by comparing what one is looking at versus what one is looking for¹⁰.

How are these goals set? The PPC integrates past visuomotor experience and rewards¹⁶, establishing bi-directional links between attention and learning¹⁵. Visual stimuli repeatedly paired with rewards gradually gain priority on the PPC map and will be preferred in visuomotor interactions such as grasping actions. When a primate faces an array of potentially valuable options, PPC subpopulations representing them compete for behavioral selection⁹. Yet, we do not understand the mechanisms that enable the rapid integration of reinforcement into PPC maps. This is in part because most studies of reward learning employed a handful of spatially unstructured options that may require little involvement of the PPC.

Here, we considered how encoding of reinforcement in the PPC, occipito-temporal and prefrontal DAN regions may resolve option competition and enable exploitation. We experimentally manipulated the distribution of rewards during rapid movement through a one-dimensional continuous space, as a clock hand revolves around a circle (Figure 1A), inducing value-laden continuous visuomotor representations. We tested the general hypothesis that representations of reinforcement history in PPC are integrated into a map that supports exploitation of the most-

rewarded option. Given the rapid dynamics of visuomotor learning in PPC, prior studies have considered working memory (WM)^{17,18} or serial hypothesis testing^{14,19}. Leveraging a previously validated computational model²⁰, we demonstrate, however, that option competition in the DAN cannot be fully explained by WM or traditional RL, but involves an information-compressing RL process that selectively maintains the values of preferred options and allows non-preferred alternatives to decay.

The key insight here is that the entropy of values within a map tunes the explore-exploit balance²⁰ (cf. policy entropy in artificial intelligence²¹). Consider a learning agent who tracks the expected reward or value associated with each target or option, termed the value function. When the estimated values of all options are equal, the entropy of the value function is highest, and the agent can needs to explore to discovery truly superior options. Conversely, when a single superior option (global value maximum) can be exploited, entropy is low. Thus, entropy of the value function quantifies global uncertainty about which option is best. Exploration generally increases the mutual information between the learned (agent's) and objective (environment's) value functions. However, a reward-maximizing agent only needs to discover the highest-valued options, rather than attempting to learn precisely the value of every option²². Furthermore, maintaining and updating a detailed map incurs a high memory cost and a risk of cognitive failure, which humans strive to minimize²³. Thus, a resource-rational agent should reduce the entropy of its learned value function^{cf. 24}.

We have previously shown that selectively maintaining the values of preferred options and forgetting non-preferred alternatives reduces the entropy or compresses the information contained in the value function²⁰. This information-compressing RL model learns and forgets faster than traditional RL, almost as guickly as classical buffer WM models. Yet, while working memory excels in recognizing deterministic rules, its limited buffer becomes a hindrance in stochastic environments. By contrast, resource-rational information-compressing RL integrates reinforcement over a period sufficiently long to explore and exploit stochastic environments efficiently. We have previously shown in a continuous space that it outperforms more memory-intensive traditional RL with long-term value persistence²⁰. Moreover, whereas WM models fail to explain empirical findings of learned long-term value or salience signals in the PPC^{25, 26}, information-compressing RL accounts for them and makes a key neural prediction: Increases in the entropy of the learned value function, and consequently the number of potentially good options, should recruit more PPC neuronal subpopulations representing them. Entropy decreases should have the opposite effect, as fewer subpopulations dominate the output and behavior shifts from exploration to exploitation. Critically, traditional RL does not predict entropy decreases during successful learning and does not link entropy dynamics to exploitation²⁰. By contrast, WM models predict divergent entropy dynamics, determined only by the content of the buffer.

Here, we adjudicate among alternative accounts of value-based option competition in the DAN: traditional RL, information-compressing RL, and WM, alone or in combination. Biophysical models and electrophysiological studies of option competition in the posterior cortex suggest that each competing option may be encoded by a subpopulation with a unique phase of β_1/α oscillatory output^{17,18,27–3°}. Thus, we hypothesized that increases in the number of close-valued options would

induce a β_1/α desynchronization. We also examined whether BOLD and oscillatory dynamics consistent with information-compressing RL would predict a successful explore-exploit transition. Two studies of DAN BOLD and one MEG study of posterior oscillations provided evidence supporting information-compressing RL.

Results

We begin by describing (i) behavior on the clock task and our information-compressing RL model, SCEPTIC (StrategiC Exploration/exploiTation of Instrumental Contingencies) and (ii) the connectivitybased parcellation of the human DAN used here. Our main analyses focus on entropy dynamics and the transition from exploration to exploitation. We report on distinct neural substrates of exploration elsewhere.



Figure 1. Paradigm and SCEPTIC model

(A) The clock paradigm consists of decision and feedback phases. During the decision phase, a dot revolves 360° around a central stimulus over the course of four seconds. Participants press a button to stop the revolution and receive a probabilistic outcome.

(B) Evolution of subjects' response times (RT) by contingency and performance. Panels represent participants whose total earnings were above or below the sample median.

(C) Rewards are drawn from one of four monotonically time-varying contingencies: two learnable (increasing expected value, IEV, and decreasing expected value, DEV, dark colors), and two unlearnable (constant expected value, CEV, and constant expected value–reversed, CEVR, light colors). Reward probabilities and magnitudes vary independently.

(D) Evolution of subjects' response time swings (RT swings) by contingency and performance.

(E) SCEPTIC model: basis function representation. Top: Subject responds at 1s and wins 110 points. Bottom left: The 1D space of the task is tiled with Gaussian-shaped learning elements with staggered receptive fields. Bottom right: the reward at 1s updates expected values (weights) of nearby basis elements.

(F) Entropy dynamics of the information-compressing model, mechanism. Left: example value distribution early in learning, when all locations have similar values and entropy is high (top: visual space of the task, bottom: linear coordinates with location on abscissa. Right: example value distribution late in learning, contrasting information-compressing RL (black line, colored bases) with traditional RL (grey line). Information compression (green arrows) reduces the entropy of the value distribution.

(G) Entropy dynamics of the information-compressing vs. traditional RL model across the entire experiment. High initial entropy reflects random uniform prior basis element weights, however qualitatively the same dynamics are seen with priors of o (Figure s1D-F).

(H) The same as G, for an average run. First run is excluded to eliminate effect of priors.

Behavior and SCEPTIC model

On the clock task (Figure 1A), participants explore and learn reward contingencies within a foursecond time interval, presenting a challenging unidimensional continuous environment. The passage of time is marked by the rotation of a dot around a clock face, reducing demands on internal timing. They are told to find the response time that yields the most points. To encourage extensive exploration and trial-by-trial learning, the task employs four stochastic reward contingencies with varying reward probability/magnitude tradeoffs (Figure 1C), which require integration of reinforcement over time and impede purely WM-based or heuristic strategies. Indeed, whereas people's responses shifted toward value maxima in learnable contingencies (Figure 1B), even more successful participants tended not to respond as early as possible in DEV or as late as possible in IEV. Thus, participants generally did not recognize that contingencies were monotonic, instead searching for a subjective value maximum (RT_{Vmax}); their estimate of its location often shifted within the block. Trial-wise changes of response times (aka 'RT swings') provide a model-free index of exploration. Early in learning, better-performing participants displayed large RT swings followed by a decline as they shifted to exploiting the subjective value maximum. Less successful participants kept exploring stochastically, with moderately large RT swings throughout, never settling on a clear value maximum. Curiously, successful participants transitioned from exploration to exploitation even in unlearnable contingencies where no objective value maximum exists²⁰. As detailed in the next section, this behavior can be explained by adaptive selective maintenance of reinforcement histories.

Our SCEPTIC reinforcement learning model²⁰ quantifies both local reinforcement (reward prediction errors) and global value map updates. On the clock task SCEPTIC approximates the value function or expected reward across the space (interval) with a set of learning elements whose temporal receptive fields cover the interval^{31,32}. Each element learns from temporally proximal rewards, updating its weight by reward prediction errors or the discrepancy between model-predicted reward at the chosen

RT and the obtained reward (Figure 1E). The highest-valued RT is the global maximum (aka RT_{Vmax}) of the model-estimated value function (Figure 1F).

To understand how global map updates can be quantified, we can think of locations (basis elements) as an alphabet, reinforcement sequences as messages, and the value function as an information source encoding the reinforcement history. The information content of this source is given by Shannon's entropy of the value function (normalized element weights), which is high when multiple attractive options compete and low when a single option dominates (Figure 1F). Thus, increases in entropy reflect the emergence of competing options on the global map. We have previously found that human behavior on the clock task is best explained by a model that selectively maintains the values of favored actions and allows the alternatives to decay, compressing the information content – that is, decreasing the entropy – of the value function²⁰. This compression heightens the relative dominance of the best actions and thus facilitates exploitation and efficient exploration (Figure 1G-H, s1D-E). These information dynamics scale with performance and non-verbal intelligence²⁰. To test the neural predictions of this model, we examine whether activity in the DAN is more consistent with an information-compressed value map or that from an otherwise identical traditional RL variant of SCEPTIC with long-term persistence of values (for details, see Methods).

Connectivity-based parcellation of the human DAN



Figure 2. Human dorsal attention network (DAN) and responses to value entropy and its change.

(A) DAN nodes arranged along the visuomotor transformation gradient, connectivity-based parcellation of Schaefer et al. (2018), details: Table s1. Detailed parcellation: Figure s2.

(B) Responses to value entropy (left) and entropy change (middle), voxel-wise GLM; Schaefer DAN parcellation for the same axial slice (right; z = 55). Responses to reward/omission: Figure s3.

Modern studies of functional brain connectivity in the human cortex reliably identify a dorsal attention network³³⁻³⁶, encompassing the temporo-occipital (putative human MT+), posterior parietal (IPS, SPL) and frontal premotor regions (FEF, PMv, PMd). A prominent parcellation

of human functional network structure^{33,34} further subdivides the DAN into two subnetworks along the caudo-rostral visuomotor gradient: the caudal subnetwork, consisting of MT+ and caudal PPC, and the rostral subnetwork consisting of rostral PPC and frontal premotor regions (Figure 2). Below, we describe activity across these four groups of regions. Since DAN subregions are characterized extensively in the macaque, for reference we label the human connectivity-based subregions according to their putative homology with monkey areas^{37–44} (Figure s2, Table s1; "putative human" is omitted from region names below for simplicity since the interpretation of our results does not depend on precise homology between human and monkey areas).

Value information dynamics in the dorsal stream and the transition from exploration to exploitation

While fMRI and MEG cannot access option representations in individual neurons or subpopulations, our analyses of learned value information dynamics enabled us to adjudicate among competing accounts based on population-level measures. Specifically, if a subpopulation is recruited to represent the value of each option, the number of active subpopulations should scale with the information content of the learned value function. Thus, we can test whether maps contained in the human DAN undergo reinforcement-based updates as predicted by our model by regressing trial-by-trial entropy change against BOLD signal and posterior oscillatory power (Figure 1F). Entropy change reflects a global update to the dispersion of values for chosen and unchosen options and is distinct from prediction errors, which only reflect the local update to the chosen option, conceptually as well as statistically (|r|<0.1 for signed or absolute prediction errors across both samples reported here). Moreover, the information-compressing vs. traditional RL variants of the SCEPTIC model make different predictions about the nature of entropy change. Under the information-compressing model, entropy change has two components: 1) the decay of unchosen options, which reduces entropy of the value function and 2) value updates to the chosen action, which can increase or decrease entropy depending on whether the update promotes the dominant option relative to alternatives. By comparison, entropy change under the traditional RL model depends only on updates to the chosen action and entropy is generally higher relative to the information-compressing model²⁰. To demonstrate information-compressing learning dynamics, we contrasted the neural fit of our information-compressing model with an otherwise identical traditional RL comparator lacking information compression. We further ascertained that observations supporting it are not explained by random between-persons heterogeneity or confounds through fine-grained analyses of within-trial activity, stringent type I error control, sensitivity analyses, behavioral validation, and out-of-session and out-of-sample replication.

DAN BOLD scales with model-predicted value map information dynamics

Our whole-brain analysis revealed that the number of potentially advantageous options measured by model-predicted value entropy and its change (Figure 2B; Tables s2-s3)¹ recruited frontoparietal regions of the DAN, but not MT+, the basal ganglia or thalamus. Entropy change additionally recruited nodes of the cinguloopercular network (dorsal ACC and anterior insula/frontal operculum) and the rostrolateral prefrontal cortex.

Frontoparietal DAN nodes but not MT+ specifically track entropy change and not novelty

If DAN responses correlated with entropy change reflect value map updates, then activity should be modulated post-reinforcement. Indeed, analyses of within-trial BOLD activity revealed that entropy change modulated frontoparietal DAN activity post-outcome, particularly in caudal PPC and frontal-premotor nodes (Figure 3A). Responses in MT+ were much weaker, in contrast to responses to scalar value of the best option (Figure 3C), which were positive in MT+ and negative in fronto-parietal nodes. Effects of entropy change were evident with or without accounting for between-subject heterogeneity (individual random slopes), behavioral confounds (current and lagged response times) and spatially non-specific reinforcement features (scalar V_{max} [Figure 3C], reward/omission, prediction error all included as covariates in Figure 3A; model without covariates: Figure s4 A).

Value entropy generally rises during early exploration when many options are sampled (Figure 1H), and one can argue that its association with neural responses is an artifact of novelty. To rule out this confound, we manipulated value entropy by changing the reward contingency every 40 trials without any explicit cues in a follow-up study of 142 older individuals with and without psychopathology. In this replication study, frontoparietal DAN responses to entropy change were qualitatively unchanged, even though we had excluded the first 10 trials from analyses to eliminate novelty effects (Figure 3B).

¹ We note that including entropy, entropy change, and prediction errors simultaneously in GLM analyses does not meaningfully change the pattern of results. This is due to the relatively low level of correlation among these signals.



Figure 3. Information dynamics of the value function, DAN BOLD signal.

(A) Responses to value entropy change (higher signal to increases reflecting a rising number of potentially valuable options), multi-level analysis of deconvolved BOLD signal from the original study, TR = 1s.

(B) Same, replication sample of older adults with and without depression, TR = 0.6s. NB: Since BOLD response is smooth, we can only interpret the peak response as indicating the timing of activity.

(C) Response to scalar Vmax, original study. D. Neural model comparison providing evidence of information-compressing rather than traditional reinforcement learning (RL). E. Analogous comparison demonstrating that DAN responses to entropy change cannot be explained solely in terms of spatial working memory updates.

Frontoparietal BOLD dynamics specifically support reinforcement learning with information compression

A hallmark of traditional instrumental learning is the long-term persistence of option values. In contrast, in our information-compressing model of learning, values of preferred regions are maintained, whereas values of spatiotemporally distant alternatives decay. To find evidence of such compression, we compared the neural fit of entropy change signals from the information-compressing model vs. an otherwise identical model without compression. The information-compressing model better accounted for DAN responses to entropy change, particularly in PPC-caudal and frontal-premotor regions, but not in MT+ (AlC_{selective} – AlC_{full}: \geq -847, Figure 3D). Compression in the SCEPTIC model occurs as part of the reinforcement-driven update (Eq. 7), and as expected, its advantage peaks at reinforcement.

Information-compressing learning is not explained by working memory updates, but complements them

One alternative possibility is that participants perform the clock task by holding recent choices and outcomes in working memory and repeating recently rewarded choices. Human choices, however, are not adequately explained by such a process. The SCEPTIC-derived RT_{Vmax} explained substantial variance in choices even after accounting for a 5-trial buffer of choices and outcomes (fMRI session: t = 5.42, MEG session: t = 6.43, Table s4; outcomes >4 trials back had no detectable impact on choice).

Neural responses to the number of valuable options could also reflect updates to a spatial working memory buffer. Although the above behavioral analyses speak to the contrary, it was important to rule out this alternative account using neural data. By definition, spatial working memory contains the history of chosen locations (response times) and corresponding outcomes. Thus, to model working memory information content in a manner directly comparable with that of SCEPTIC, we encoded the selection history using the same representational structure (basis functions and spatial generalization gradient), adding a buffer of recent outcomes (detailed in Methods). We then predicted neural activity with the entropy and entropy change of the selection history and outcome history (reflecting working memory buffer updates, Figure s5 B-C) with and without corresponding SCEPTIC signals. As in our main analysis, responses to value entropy change peaked ~1s post-outcome (Figure s5 A), and model fit improved by \geq -182 AIC points after adding SCEPTIC predictors (Figure 3E; and after accounting for random slopes of all entropy variables, \geq -1726 points). Overall, while our analyses replicate common findings of spatial working memory representations in the DAN and specifically PPC, they support parallel information-compressing updates of option values.

Another confound related to selection history is the potential impact on value entropy and its neural correlates of preceding RT swings, whether they reflect strategic exploration or stochastic or even off-task responses⁴⁵. To rule out this possibility, we quantified a shift in the local distribution of choices as the summed Kullback-Leibler divergence (KLD; a metric of divergence between distributions) of response times for trials t-4, t-3, and t-2 from the local distribution of response times of the preceding three trials. Higher values of this measure reflect a history of larger RT swings. With (Figure 3A) or without (Figure 55A) this KLD measure as a covariate the entropy change effects were qualitatively unchanged, corroborating the notion that entropy change reflects value map updates rather than selection history. Interestingly, on trials following larger RT swings, we observed lower online rostral PPC activity and weaker frontoparietal responses to feedback (Figure 55B), potentially indicating lapses in sensorimotor activity and attention (see also a similar analysis of MEG below).

DAN sensitivity to the number of potentially valuable options predicts exploitation

The analyses above suggest that the DAN contains information-compressed value maps, but do these maps indeed govern the transition from exploration to exploitation? To answer this question, we tested whether individuals whose DAN activity better tracked with model-predicted value map updates (entropy change) made more value-sensitive, exploitative choices. We extracted entropy change regression coefficients ("betas") for each DAN parcel from individual subjects' whole-brain

analyses and entered these as between-subjects predictors in a multilevel survival model (with timepoints nested within trials and trials nested within subjects) predicting the momentary rate (hazard) of response with SCEPTIC-derived within-trial momentary value and its interaction with the fMRI beta. This interaction was positive across the parcels, indicating that individual value sensitivity scaled with entropy change responses across the DAN. This effect was replicated out-of-session (Figure 4A; anatomical distribution of behavioral effect was preserved out of session, Figure 4D) and persisted in sensitivity analyses controlling for the non-decision time (censoring the first 1s of the interval) and avoidance of missing the response window (censoring the last 0.5s; Figure s6E-F). With individual random slopes, effects were similar in the original sample, surviving FDR correction in the out-of-session replication only in premotor parcels (Figure s6A-B). To ensure that this effect did not depend on the Cox model proportional hazards assumption, we tested it in an independent trial-level GLM predicting response times with SCEPTIC-derived RT_{Vmax} the location of the highest-valued option, which enabled us to account for additional behavioral confounds (Methods, fMRI Analyses) and between-subject heterogeneity in value sensitivity (random value slopes). The results were qualitatively unchanged (Figures 4B-C, s6).



A. Effect of neural response to entropy change on exploitation, survival analysis

Figure 4. BOLD encoding of value information dynamics and behavioral exploitation.

(A) Multi-level survival analyses examining how the individual's neural response moderates their behavioral sensitivity to within-trial time-varying value. Left: original fMRI session. Right: replication, MEG session. Greater modulation of individual DAN BOLD response by value entropy change predicted more exploitative choices.

(B), (C) Same, GLM analysis.

(D) The anatomical pattern of brain-behavior associations was preserved across the original fMRI and replication sessions. Each dot represents a single DAN parcel as labeled in panel A.

Posterior β_{l}/α suppression reflects value information dynamics

Having observed dynamic value maps encoded in fronto-parietal DAN BOLD, we sought to understand cortical oscillation dynamics that underlie them. A late (550-1000ms post-feedback) $\beta_1/\alpha_$ band response has been reported on the clock task⁴⁶, but its functional significance was unclear. We hypothesized that this response reflected an update to the parietal value map, with increases in the number of valuable options resulting in global desynchronization. Indeed, increases in entropy (and the number of valuable options) elicited suppression in the 7-17 Hz (β_1/α) band at 400-750 ms, prominent in the posterior sensors (Figure 5 A, C). The reconstructed sources of this signal followed an anatomical distribution similar to the pattern observed in fMRI (Figure 5D vs. 5E). As in our analyses of BOLD, late β_1/α suppression was not explained by behavioral confounds (reward, RT_t, RT_{t-1}, V_{max}; Methods, MEG Analyses) and was strongly related to entropy change and absolute prediction errors, but not to reward/omission (Figure 5F), suggesting that β_1/α oscillations encode updates to the entire map of chosen and unchosen options, with the chosen option commanding additional processing. This β_1/α response was evident in two learnable and one unlearnable condition. However, it was almost abolished in CEVR ($\chi^2(3) = 10.14$, p < .018) where the probability/magnitude trade-off was the opposite of other conditions, indicating that the response was altered when outcomes did not match one's expectations based on experience with previously encountered environments. This late suppression spread into the theta band, peaking at 600-800 ms and 3-6 Hz, evident mostly in posterior sensors. Additionally, an earlier burst of suppression at 8-17 Hz emerged immediately following response and ceased after the outcome (Figure 5A), suggesting that participants were at times anticipating an increase in global uncertainty based on their response.

It is possible that, instead of β_1/α desynchronization to entropy increases, our observations reflect β_1/α synchronization to entropy *decreases* relative to baseline. We ruled out this possibility by separately examining the effects of entropy increase (vs. decrease or no change) and entropy decrease (vs. increase or no change; Figure s7A-B). Whereas entropy increases elicited massive suppression at 8-20 Hz peaking at 0.4-0.8s post-outcome and spreading into the theta band, entropy decreases did not elicit synchronization of a similar magnitude.

One could also argue that effects of entropy increases merely reflect a recent history of highly variable choices⁴⁵ rather than updates to the distribution of learned value across competing options. Interestingly, while a recent history of RT swings measured by the Kullback-Leibler distance between $RT_{\{t-3, t-2\}}$ and RT_{t-1} , predicted suppression in the 7-16 Hz band (same model as above, Figure s7C),

effects of entropy increases persisted while controlling for RT swings (Figure s7A), indicating that both selection and reinforcement history are encoded in β_1/α oscillations.

Finally, manipulation effects are heterogeneous across individuals⁴⁷, and we verified that our findings were reliable after accounting for inter-individual heterogeneity by including the subject random slope of entropy change in our multi-level models (Figure s7D).

Posterior β_{I}/α oscillation dynamics support information compression



Figure 5. MEG: oscillatory responses to value information dynamics and exploitation, anatomical and functional

A. Modulation of spectral power by entropy increases

relationship to BOLD signal.

(A) Oscillatory response to value entropy change: cool colors represent de-synchronization to increases reflecting a rising number of potentially valuable options, most prominent between 7-17 Hz (β_1/α) band at 400-750 ms.

(B) Neural model comparison providing evidence of information-compressing rather than traditional RL (c.f. Figure 3D). Hot colors represent AIC difference favoring the information-compressing (selective maintenance) model.

- (C) β_1/α de-synchronization was most evident in posterior sensors, consistent with a parietal source.
- (D) Source reconstruction localizes the β_1/α suppression to the posterior parietal cortex.

(E) fMRI BOLD map shown for comparison.

(F) β_1/α de-synchronization was much better explained by value entropy change or absolute reward prediction errors than by reward/omission, model controlling for individual random slopes. More negative t-statistics indicate a stronger effect.

(G) Condition-level relationships between individuals' BOLD and oscillatory responses to entropy change. Light blue bars represent individuals with stronger oscillatory responses (greater suppression to entropy increases). Y-axis: higher coefficient values indicate stronger BOLD response. X-axis: DAN regions from which BOLD response was extracted.

Our analyses of BOLD indicated that reinforcement representations in the fronto-parietal DAN nodes were compressed as predicted by SCEPTIC. To understand whether similar information-compressing dynamics were reflected in oscillatory activity, we compared the fit of the MLM with entropy change regressor derived from either the information-compressing (exactly as in our main analysis above) or the traditional RL SCEPTIC model. Indeed, the information-compressing model dominated in the in the 8-17 Hz band at 400-750 ms, in the 3-6 Hz band at 600-800 ms and at 8-20 Hz peri-response (Figure 5 B), indicating that representations of competing options reflected in oscillatory activity displayed information-compressing dynamics predicted by the SCEPTIC model.

Posterior $eta_{\prime\prime} lpha$ oscillation dynamics predict the explore-exploit transition

To understand whether β_1/α suppression responses to an increased number of options (entropy change) scaled with exploitation we used models similar to those employed in fMRI analyses (Figure 4). To ensure that our results generalized across contingencies, we decomposed summary β_1/α suppression responses into person-level means and condition-wise deviations. Person-level responses predicted exploitation (momentary value * b suppression response: z = 9.47, $\chi^2(1) = 89.69$, $p < 10^{-15}$). This effect was robust to between-subject heterogeneity (random slope of value, fixed effect: z = 2.14, $\chi^2(1) = 4.57$, p = 0.0326) and replicated out-of-session (z = 13.03, $\chi^2(1) = 169.88$, $p < 10^{-15}$), even after accounting for between-subject variability in the effect of value (z = 3.40, $\chi^2(1) = 11.60$, p < 0.001). After accounting for subject-level responses, no additional effect was observed, at the within-person, condition level ($|z| \le 0.94$, $\chi^2(1) \le 0.86$, p > 0.35), suggesting that the relationship of oscillatory responses and behavioral exploitation manifests at the between-person level.

Magnitude of posterior $\beta_{l'}/\alpha$ response scales negatively with BOLD, but only in learnable conditions

Condition-level β_1/α synchrony scaled negatively with BOLD responses across DAN, but only in learnable conditions, while the opposite pattern was seen in unlearnable conditions (Figure 5G, β_1/α response main effect: $\chi^2(1) = 10.65$, p = 0.0011, β_1/α response *condition $\chi^2(3) = 36.49$, p < 10⁻⁷), suggesting that β_1/α suppression and/or BOLD are differentially sensitive to the presence of an objective value maximum, and potentially also to the match between current and previously encountered contingencies.

In summary, we found that the fronto-parietal nodes of the dorsal stream represented a compressed reinforcement history, mapping values of potentially valuable options as did posterior β_1/α oscillations. These neural dynamics predicted a successful transition from exploration to exploitation.

Discussion

When exploring and exploiting a few discrete options, primates rely on choice and reinforcement histories encoded in the striatum and amygdala. Additional demands, however, arise when moving rapidly through space and choosing when to harvest a reward. Our multimodal imaging study of human BOLD signal and cortical oscillations revealed that reward-based learning in such an environment involves dynamic value maps in the dorsal stream. More specifically, we found that BOLD signals in the PPC and premotor cortex increased and posterior β_1/α oscillations desynchronized in response to increases in the number of valuable options and, correspondingly, uncertainty about the best option. This global uncertainty was quantified by changes in the entropy of the learned value function captured by our computational model. These BOLD and oscillatory dynamics predicted a successful behavioral transition from exploration to exploitation, with out-of-session and out-of-sample replication. BOLD dynamics consistent with map updates were seen throughout the parietal and frontal nodes of the dorsal stream, but not in the occipito-temporal MT+.

Much debate about maps in the dorsal stream, especially in the PPC, has focused on what they represent. They have been suggested to encode attentional priority⁴⁸⁻⁵¹, the intention to move⁵², expected value vs. salience of stimuli^{25,26,53}, or expected information gain^{54,55}. Disagreements between studies are not entirely explained by anatomical heterogeneity or methodological differences¹⁰. Rather, this debate may be resolved in part by recognizing that real-world motor programs are inextricable from visuospatial and value-laden representations of targets. This aligns with the affordance perspective in which sensorimotor systems continuously encode opportunities for action emerging in the immediate environment^{9,56}. Affordance representations throughout the dorsal stream multiplex visual, oculomotor, motor, and somatosensory information^{57–59}. We observed that the nodes in the dorsal stream along the visuomotor gradient from caudal PPC to premotor cortex respond similarly to the values of options, consistent with the notion of pragmatic, multimodal affordance representations. These dynamics were considerably weaker in the MT+, which instead responded to the recent reward and long-term value, potentially indicating that processing of visual motion (here, ball motion around the clock face) was enhanced when the expected reward rate was high, as electrophysiological studies suggest⁶⁰. Thus, fronto-parietal regions but not MT+ contain a spatially structured value map for all options.

Our results address the critical question of how competition between multiple affordances is resolved^{9,61,62}. Cisek and colleagues speculate that the prevailing affordance in output regions is determined by both reinforcement learning and goals signaled to the dorsal stream by ventral prefrontal systems^{63–65}. We find that chosen and unchosen option values are updated on the PPC map within 0.4–0.7s of reinforcement, predicting exploitation of the highest-valued option. Crucially, dorsal stream value map updates and behavior were better described by an information-compressing reinforcement learning algorithm relative to traditional instrumental learning, even after accounting for updates to the visuospatial WM buffer⁶⁶. This algorithm selectively maintains preferred actions, compressing information about learned values and supporting more efficient exploitation during continuous visuomotor interactions.

Our observations are not easily explained by an earlier account of reward learning in discrete spaces postulating that it relies on early (0.2-0.4s) traditional RL updates and later (0.4-0.7s) working memory updates⁶⁷. During continuous sensorimotor interaction we observe both value map and spatial working memory buffer updates 0.4-0.7s post-outcome. How does the PPC obtain information about incoming reinforcement? The identity of a reward or goal state may be cached in the dorsal and ventral stream⁶⁵. Thus, upon reaching the goal, reinforcement may be almost instantaneous⁶³. Reinforcement may also be signaled meso-striato-thalamo-cortical reward prediction errors. However, these signals are unlikely to arrive early enough to enable the observed value map updates⁶³ and thus may shape learning only at slower timescales. In any case, since updates to values of *unchosen* regions of the continuous space must account for the spatial proximity of the sampled point, they can only be updated by a network that contains a full map, such as the PPC. Our results suggest that a dual-systems account of competing frontoparietal WM and meso-striatal RL controllers⁶⁷ may not extend to the rapid continuous sensorimotor interaction. Affordance competition must in part be resolved through reinforcement learning in the dorsal stream, with information compression facilitating a "within-system" decision⁶⁸.

The connection between information-compressing RL and β_1/α oscillations bridges our populationlevel account of option competition in the dorsal stream with biophysically realistic circuit models. Gelastopoulos, Whittington and Kopell¹⁸ propose that competing representations in the parietal cortex are carried by β_1/α -synchronized ensembles organized along cortical columns. It is thought that recurrent excitation stabilizes preferred options, while lateral inhibition suppresses non-preferred alternatives^{18,69,70}. In line with these circuit-level models and empirical studies, we propose that during the value-guided exploration of the sensorimotor space, recruitment of many ensembles produces an asynchronous oscillatory output, reflecting greater entropy of the value map and global uncertainty about the best action (Figure 6). When an option is preferentially sampled and reinforced, the regional output becomes dominated by the β_1/α -synchronized ensemble representing this option, promoting exploitation. Compression of the value function may depend on lateral inhibition in which dominant ensembles suppress and even highjack the β_1/α output of competing counterparts, reducing their likelihood of behavioral selection^{18,69}.



 \downarrow BOLD, β_1/α synchronization

Figure 6. Option competition in the posterior parietal cortex: conceptual model.

Top: exploration mode. Top left: Multiple options compete for selection and the entropy of the value function is high. Top right: Competing options are represented by neuronal subpopulations with each producing oscillatory output with a distinct phase.

Bottom: exploitation mode. Top left: following information-compressing learning, a global value maximum emerges. Top right: as a result of recurrent excitation and lateral inhibition, the subpopulation representing the dominant option begins to dominate the output. After Gelastopoulos, Whittington and Koppel (2019; biophysical model of β_1/α -stabilized competing cortical populations) and Mysore and Kothari (2020; computational models of competitive selection).

We observe a clear functional correspondence between dorsal stream BOLD and posterior β_1/α desynchronization: value entropy change modulated BOLD positively (Figure 3A-B, 4B-c) and β_1/α power negatively (Figure 5A, E). Interestingly, BOLD and β_1/α responses were correlated only in learnable contingencies. Moreover, oscillations were sensitive to a mismatch with previous experience: in a contingency with an unexpectedly reversed probability/magnitude tradeoff, oscillatory responses no longer tracked entropy dynamics. While the spatial resolution of fMRI complements the temporal resolution of electrophysiology, BOLD and oscillatory power capture different aspects of cortical local field potentials. Their empirical correlations are generally positive in gamma band and negative in α and β (which contain additional unique information about BOLD), with β but not α suppression accelerating increases and delaying decreases in BOLD^{71–73}. Thus, our fMRI and MEG findings align and ostensibly reflect updates to the dynamic value map. These functional properties were not shared by other frequency bands such as high beta, theta and delta.

The strengths of our study include consistent findings of value entropy dynamics in the human dorsal stream across MEG and fMRI modalities, across sessions, and in a separate fMRI sample. Out-of-session and out-of-sample replications increase confidence in the observed links between dynamic maps and behavioral exploration/exploitation. Our computational model comparisons supported the conclusion that value maps in the dorsal stream are likely shaped by an information-compressing RL process that cannot be explained by traditional instrumental learning or WM buffer accounts. Experimental manipulation of reinforcement in a continuous space and our RL model provided access to a spatially structured value vector, dissociating global from local updates. Our novel multilevel analyses revealed parallel within-trial temporal dynamics of cortical oscillations and deconvolved BOLD signals. Finally, our observations of value entropy dynamics replicated in a modified experiment with unsignaled reversals, ruling out novelty as an alternative explanation.

The main limitation of our study is the lack of a causal manipulation that would isolate contributions of various DAN nodes to resolving the explore-exploit dilemma, although our findings of dynamic value maps in the dorsal stream are in line with human transcranial stimulation studies^{53,74}. Human neural stimulation and rodent optogenetic studies are needed to test the model of value-dependent affordance competition articulated here. It is also difficult to know to what extent our findings generalize to non-temporal spaces and to punishments as opposed to rewards. Finally, MEG and fMRI data were collected in separate sessions, enabling out-of-session replication, but precluding an analysis of simultaneous BOLD and cortical oscillation recordings.

In conclusion, exploration and exploitation of a continuous sensorimotor space depend on dynamic value maps in the dorsal stream, particularly in the PPC. Our observations suggest that the dorsal stream selectively maintains values of preferred options and compresses out inferior, spatiotemporally distant alternatives. Indeed, as circuit models of option competition in the dorsal cortex suggest, β_1/α oscillatory output of posterior cortical subpopulations desynchronizes when more competing options emerge and synchronizes when non-preferred alternatives are compressed out. Compression notwithstanding, we show that option values in PPC persist beyond the timescale of the WM buffer. Our results support the affordance competition view of maps in dorsal cortex and are at odds with the notion that sensorimotor choices require a sequence of temporally distinct sensory, reward, cognitive and motor computations. Altogether, our study sheds light on how primates, including humans, track the values of alternative options in complex, rapidly changing environments.

Methods

Participants

Participants in the original study were 70 typically developing adolescents and young adults aged 14– 30 (M = 21.4, SD = 5.1). Thirty-seven (52.8%) participants were female and 33 were male. Prior to enrollment, participants were interviewed to verify that they had no history of neurological disorder, brain injury, pervasive developmental disorder, or psychiatric disorder (in self or first-degree relatives). Participants in the replication study were 143 middle-aged and older adults aged 50-80 (*M* = 62.2, *SD* = 6.8), 80 (56%) were female and 62 were male; 101 were diagnosed with DSM-IV non-psychotic major depression. Individuals with a history of psychosis, mania, neurological conditions of the brain and current substance use disorders were excluded from the replication study. Participants and/or their legal guardians provided informed consent or assent prior to participation in both studies. Experimental procedures for this study complied with Code of Ethics of the World Medical Association (1964 Declaration of Helsinki) and the Institutional Review Board at the University of Pittsburgh (protocols PRO10090478 and STUDY19030288). Participants were compensated \$75 for completing the original experiment and \$150 for completing the replication study, which included other experiments.

Procedure

Original study

As part of a larger study, participants completed an exploration and learning task ("clock task"; Figure 1A, and detailed below) in separate magnetoencephalography (MEG) and functional MRI (fMRI) sessions. The order of the fMRI and MEG sessions was counterbalanced (fMRI first n = 34, MEG first n = 36) and the sessions were separated by 3.71 weeks on average (SD = 1.59 weeks).

During the fMRI session, participants completed eight runs of the clock task (based on Moustafa et al., 2008). Runs consisted of 50 trials in which a green dot revolved 360° around a central stimulus over the course of 4s. Participants pressed a button to stop the dot, which ended the trial. They then received a probabilistic reward for the chosen response time (RT) according to one of four time-varying contingencies, two learnable (increasing and decreasing expected value) and two unlearnable. All contingencies were monotonic but featured reward probability/magnitude tradeoffs that made learning difficult (see²⁰ for more detailed analyses of the task). After each response, participants saw the probabilistic reward feedback for 0.9s. If participants failed to response within 4s, they received zero points.

The central stimulus was a face with a happy expression or fearful expression, or a phase- scrambled version of face images intended to produce an abstract visual stimulus with equal luminance and coloration. Faces were selected from the NimStim database⁷⁵. All four contingencies were collected with scrambled images, whereas only IEV and DEV were also collected with happy and fearful faces. The effects of the emotion manipulation will be reported in a separate manuscript because they are not central for the examination of the neural substrates of exploration and exploitation on this task.

Each trial was followed by an intertrial interval (ITI) that varied in length according to an exponential distribution. To maximize fMRI detection power, the sequence and distribution ITIs were derived using a Monte Carlo approach implemented by the *optseq2* command in *FreeSurfer* 5.3. More specifically, we simulated five million possible ITI sequences consisting of 50 trials each and retained the top 320 orders based on their estimation efficiency. For each subject, the experiment software randomly sampled 8 of these efficient ITI sequences, which were used for the durations of ITIs in the task.

During the MEG session, participants completed eight runs of the same task. The contingencies and trial structure were identical to fMRI (see Figure 1A), requiring participants to respond within a four-

second interval to maximize the points they earned. Given the lower signal-to-noise ratio of MEG relative to fMRI, runs consisted of 63 trials each.

As detailed in the results, the behavioral data from the MEG and fMRI sessions were used to test the out-of-session consistency of brain-behavior effects identified by each modality. This enabled us to establish whether individual differences in dorsal stream activity and exploration/exploitation represented stable tendencies vs. patterns incidental to a single experimental session.

Replication study

Procedures of the replication study were similar, but no MEG data were collected. Participants completed 240 trials of the clock task in two runs. Only IEV and DEV contingencies were employed. To dissociate value entropy from novelty, the contingency reversed every 40 trials unbeknownst to the participants. Trials were extended to 5s to accommodate slower psychomotor speed in this older sample.

Imaging Acquisition and Processing Methods

fMRI acquisition

Neuroimaging data during the clock task were acquired in a Siemens Tim Trio 3T scanner for the original study and Siemens Tim Prisma 3T scanner for the replication study at the Magnetic Resonance Research Center, University of Pittsburgh. Due participant-dependent variation in response times on the task, each fMRI run varied in length from 3.15 to 5.87 minutes (M = 4.57 minutes, SD = 0.52). Functional imaging data for the original/replication study were acquired using a simultaneous multislice sequence sensitive to BOLD contrast, TR = 1.0/0.6s, TE = 30/27ms, flip angle = 55/45°, multiband acceleration factor = 5/5, voxel size = 2.3/3.1mm³. We also obtained a sagittal MPRAGE T1-weighted scan, voxel size = 1/1mm³, TR = 2.2/2.3s, TE = 3.58/3.35ms, GRAPPA 2/2x acceleration. The anatomical scan was used for coregistration and nonlinear transformation to functional and stereotaxic templates. We also acquired gradient echo fieldmap images (TEs = 4.93/4.47ms and 7.39/6.93ms) for each subject to mitigate inhomogeneity-related distortions in the functional MRI data.

Preprocessing of fMRI data

Anatomical scans were registered to the MNI152 template⁷⁶ using both affine (ANTS SyN) and nonlinear (FSL FNIRT) transformations. Functional images were preprocessed using tools from NiPy⁷⁷, AFNI (version 19.0.26)⁷⁸, and the FMRIB software library (FSL version 6.0.1)⁷⁹. First, slice timing and motion coregistration were performed simultaneously using a four-dimensional registration algorithm implemented in NiPy⁸⁰. Non-brain voxels were removed from functional images by masking voxels with low intensity and by the *ROBEX* brain extraction algorithm⁸¹. We reduced distortion due to susceptibility artifacts using fieldmap correction implemented in FSL FUGUE.

Participants' functional images were aligned to their anatomical scan using the white matter segmentation of each image and a boundary-based registration algorithm⁸², augmented by fieldmap unwarping coefficients. Given the low contrast between gray and white matter in echoplanar scans

with fast repetition times, we first aligned functional scans to a single-band fMRI reference image with better contrast. The reference image was acquired using the same scanning parameters, but without multiband acceleration. Functional scans were then warped into MNI152 template space (2.3mm output resolution) in one step using the concatenation of functional-reference, fieldmap unwarping, reference-structural, and structural-MNI152 transforms. Images were spatially smoothed using a 5mm full-width at half maximum (FWHM) kernel using a nonlinear smoother implemented in FSL SUSAN. To reduce head motion artifacts, we then conducted an independent component analysis for each run using FSL MELODIC. The spatiotemporal components were then passed to a classification algorithm, ICA-AROMA, validated to identify and remove motion-related artifacts⁸³. Components identified as noise were regressed out of the data using FSL regfilt (non-aggressive regression approach). ICA-AROMA has performed very well in head-to-head comparisons of alternative strategies for reducing head motion artifacts⁸⁴. We then applied a .oo8 Hz temporal high-pass filter to remove slow-frequency signal changes⁸⁵; the same filter was applied to all regressors in GLM analyses. Finally, we renormalized each voxel time series to have a mean of 100 to provide similar scaling of voxelwise regression coefficients across runs and participants.

Treatment of head motion

In addition to mitigating head motion-related artifacts using ICA-AROMA, we excluded runs in which more than 10% of volumes had a framewise displacement (FD) of 0.9mm or greater, as well as runs in which head movement exceeded 5mm at any point in the acquisition. This led to the exclusion of 11 runs total, yielding 549 total usable runs across participants. Furthermore, in voxelwise GLMs, we included the mean time series from deep cerebral white matter and the ventricles, as well as first derivatives of these signals, as confound regressors⁸⁴.

MEG Data acquisition

MEG data were acquired using an Elekta Neuromag VectorView MEG system (Elekta Oy, Helsinki, Finland) in a three-layer magnetically shielded room. The system comprised of 306 sensors, with 204 planar gradiometers and 102 magnetometers. In this project we only included data from the gradiometers, as data from magnetometers added noise and had a different amplitude scale. MEG data were recorded continuously with a sampling rate of 1000 Hz. We measured head position relative to the MEG sensors throughout the recording period using 4 continuous head position indicators (CHPI) that continuously emit sinusoidal signals, and head movements were corrected offline during preprocessing. To monitor saccades and eye blinks, we used two bipolar electrode pairs to record vertical and horizontal electrooculogram (EOG).

Preprocessing of MEG data

Flat or noisy channels were identified with manual inspections, and all data preprocessed using the temporal signal space separation (TSSS) method^{86,87}. TSSS suppresses environmental artifacts from outside the MEG helmet and performs head movement correction by aligning sensor-level data to a common reference⁸⁸. This realignment allowed sensor-level data to be pooled across subjects group analyses of sensor-space data. Cardiac and ocular artifacts were then removed using an independent component analysis by decomposing MEG sensor data into independent components (ICs) using the

infomax algorithm⁸⁹. Each IC was then correlated with ECG and EOG recordings, and an IC was designated as an artifact if the absolute value of the correlation was at least three standard deviations higher than the mean of all correlations. The non-artifact ICs were projected back to the sensor space to reconstruct the signals for analysis. After preprocessing, data were epoched to the onset of feedback, with a window from -0.7 to 1.0 seconds. Trials with gradiometer peak-to-peak amplitudes exceeded 3000 fT/cm were excluded. For each sensor, we computed the time-frequency decomposition of activity on each trial by convolving time-domain signals with Morlet wavelet, stepping from 2 to 40 Hz in logarithmic scale using 6 wavelet cycles. This yielded trial-level time-frequency data that were amenable to multilevel models across frequencies and peri-feedback times.

Computational Model of Behavior

Core architecture of SCEPTIC reinforcement learning (RL) model

The SCEPTIC model represents the one-dimensional space/time of the clock task using a set of unnormalized Gaussian radial basis functions (RBFs) spaced evenly over an interval *T* in which each function has a temporal receptive field with a mean and variance defining its point of maximal sensitivity and the range of times to which it is sensitive, respectively (a conceptual depiction of the model is provided in Figure 1). The primary quantity tracked by the basis is the expected value of a given choice (response time; we use the intuitive term *value* for continuity with prior studies of PPC maps ^{25, 26, 53}, however since this estimate does not converge on the true reward rate, it is technically a *preference*, see text following eq. 7). To represent time-varying value, the heights of the basis function of each basis function to the integrated value representation depends on its temporal receptive field:

$$\varphi_b(x) = \exp\left[-\frac{(x-\mu_b)^2}{2s_b^2}\right]$$
 (1)

where x is an arbitrary point within the time interval T, μ_b is the center (mean) of the RBF and s_b^2 is its variance. And more generally, the temporally varying expected value function on a trial *i* is obtained by the multiplication of the weights with the basis:

$$V(i) = \sum_{b=1}^{B} w_b(i)\varphi_b$$
⁽²⁾

For the clock task, where the probability and magnitude of rewards varied over the course of foursecond trials, we spaced the centers of 24 Gaussian RBFs evenly across the discrete interval and chose a fixed width, s_b^2 , to define the temporal variance (width) of each basis function. More specifically, s_b^2 was chosen such that the distribution of adjacent RBFs overlapped by approximately 50% (for details and consideration of alternatives, see²⁰).

The basic model, referred to as *traditional RL* in Results, learns the expected values of different response times by updating each basis function *b* according to the equation:

$$w_b(i+1) = w_b(i) + e_b(i|t)\alpha[\text{reward}(i|t) - w_b(i)]$$
 (3)

where *i* is the current trial in the task, *t* is the observed response time (aka RT), and reward(i|t)

is the reinforcement obtained on trial *i* given the choice *t*. Prediction error updates are weighted by the learning rate α and the temporal generalization function or eligibility *e*. To avoid tracking separate value estimates for each possible moment, feedback obtained at a given response time *t* is propagated to adjacent times. Thus, to implement temporal generalization of expected value updates, we used a Gaussian RBF centered on the response time *t*, having width s_g^2 . The eligibility e_b of a basis function φ_b to be updated by prediction error is defined as its overlap with the temporal generalization function, *g*:

$$g(x) = \exp\left[-\frac{(x-t)^2}{2s_g^2}\right]$$
 (4)

$$e_b(i|t) = \int_0^T \min\left[g(\tau), \varphi_b(\tau)\right] d\tau$$
(5)

where τ represents an arbitrary timepoint along the interval *T*. Thus, for each RBF *b*, a scalar eligibility e_b between zero and one represents the proportion of overlap between the temporal generalization function and the receptive field of the RBF^{9°}. In the case of complete overlap, where the response time is perfectly centered on a given basis function, e_b will reach unity, resulting a maximal weight update according to the learning rule above. Conversely, if there is no overlap between an RBF and the temporal generalization function, e_b will be zero and that RBF will receive no update. Importantly, for the eligibility to be bounded on interval [0,1], the basis functions are each normalized to have an area under the curve of unity (i.e., representing probability density). Here, we also fixed the width of the generalization function to match the basis (i.e., $s_q^2 = s_b^2$).

The SCEPTIC model selects an action based on a softmax choice rule, analogous to simpler reinforcement learning problems (e.g., two-armed bandit tasks⁹¹). For computational speed, we arbitrarily discretized the interval into 100ms time bins such that the agent selected among 40 potential responses (i.e., a multinomial representation). At trial *i* the agent chooses a response time in proportion to its expected value:

$$p[t(i+1) = j|V(i)] = \frac{\exp(V(i)_j/\beta)}{\sum_{\tau=0}^T \exp(V(i)_\tau/\beta)}$$
(6)

where *j* is a specific response time and the temperature parameter β controls value sensitivity such that choices became more stochastic and less value-sensitive at higher β values.

Information-compressing RL with selective maintenance

Importantly, as detailed previously²⁰, a model that selectively maintained frequently chosen, preferred actions far outperformed model alternative models. Specifically, basis weights corresponding to non-preferred, spatiotemporally distant actions revert toward a prior in inverse proportion to the temporal generalization function:

$$w_b(i+1) = w_b(i) + e_b(i|t)\alpha[\text{reward}(i|t) - w_b(i)] - \gamma (1 - e_b(i|t))(w_b(i) - h)$$
(7)

where γ is a selective maintenance parameter between zero and one that scales the degree of

reversion toward a point *h*, which is taken to be zero here for parsimony, but could be replaced with an alternative prior expectation. Our primary fMRI analyses used signals derived from fitting the information-compressing RL model (eq. 7) to participants' behavior, while comparisons with traditional RL used the model with the learning rule described in equation 3. Two features supported by computational studies and tests against human behavior²⁰, (i) the decay in the weights of unchosen alternatives (eq. 7) and (ii) calculation of prediction errors based on individual element weights w_b (eq. 3, 7) rather than the total value estimate V(i) (eq. 2) preclude w_b or V(i) from converging on the true reward rate. While we refer to w_b as *expected values* for continuity with previous studies of the parietal cortex^{25,26,53}, w_b are closer to *preferences* in policy gradient algorithms^{22,9:93}. Here, we focus on testing the hypothesis of information compression (eq. 8) and make no strong claims about whether representations of reinforcement in PPC constitute expected values or preferences. Although taken together with our earlier behavioral and computational results, neural model comparisons reported here can be taken to favor the preferences hypothesis, a definitive adjudication will require new experiments. Value vs. policy learning accounts are not necessarily mutually exclusive since actorcritic algorithms combine both approaches.

As detailed in the Results, we defined the information content of the learned value distribution as Shannon's entropy of the normalized basis weights (the trial index *i* is omitted for simplicity of notation):

$$H(\mathbf{w}) = -\sum_{b=1}^{B} w_b \log_{10}(w_b)$$
(8)

We further sought to examine whether entropy responses in the dorsal attention network were consistent with the information-compressing selective maintenance model. To test the specificity of the representation, we conducted analyses using entropy calculated from the information-compressing SCEPTIC selective maintenance model (equation 7) vs. entropy from a traditional RL, full-maintenance counterpart that did not decay the values of unchosen actions (equation 3; detailed model comparisons provided in²⁰).

Working memory model

We argue that information dynamics attributable to value entropy from the SCEPTIC informationcompressing model best explain the exploration-exploitation transition in behavior and updates to the value map in the DAN. Yet, one could imagine that that the DAN relies solely on a spatial working memory representation with a buffer containing locations and outcomes of recent choices. In turn, the information content of this buffer might be sufficient to explain DAN BOLD activity.

We first examined whether a WM process alone is sufficient to explain human choices without invoking information-compressing RL. We used a multi-level regression model as described below (*Brain-behavior fMRI analyses using regression coefficients from model-based fMRI GLM analyses*) to predict the participant's current RT with *k* preceding RTs representing the selection history buffer and their interactions with reward/omission representing the outcome buffer. This analysis revealed no effect of outcomes beyond 4 trials (Supplementary Table 4). To assess the effect of the more remote reinforcement history not captured by the last *k* choices and outcomes, we then tested the

incremental contribution of the RT_{Vmax} (time of the learned value maximum) derived from the SCEPTIC model.

To conduct a conclusive test of this alternative account, we created a strong working memory-only comparator model that adopted the SCEPTIC RBF representation, using the basis weights to store the buffer of recently chosen locations, alongside a separate vector of recent outcomes. Specifically, the model remembered the past *k* choices by placing a unit-height eligibility functions at these locations and taking the sum, forming a selection history function, s(x). *k* was empirically estimated as 4 using multi-level linear regression (Supplemental Table 4). In turn, this representation of selection history was combined with the RBF by multiplying the selection history function and basis, yielding working memory basis weights w_b^{WM} whose height scaled with the selection history.

$$s(x) = \sum_{j=i-5}^{i-1} \exp\left[-\frac{(x-t(j))^2}{2s_g^2}\right]$$
(9)

$$w_b^{WM}(i) = \int_0^T [s(\tau) \varphi_b(\tau)] d\tau$$
(10)

where *i* represents the current trial and t(j) is the response time on the *i*th previous trial. In turn, entropy can be calculated on the selection history basis weights in the same fashion as in the regular SCEPTIC model (Equation 8). Outcome history was simply represented by a vector **o** containing 1s for rewards and os for reward omissions in the past four trials. This implementation did not require estimating free parameters from behavior. Thus, total entropy or information content of working memory H^{WM} is the joint entropy of the selection and outcome buffers:

$$H_{total}^{WM} = H(\mathbf{w}) + H(\mathbf{o}) \tag{11}$$

Fitting of SCEPTIC model to behavioral data

SCEPTIC model parameters were fitted to individual choices using an empirical Bayesian version of the Variational Bayesian Approach⁹⁴. The empirical Bayes approach relied on a mixed-effects model in which individual-level parameters were assumed to be sampled from a normally distributed population. The group's summary statistics, in turn, were inferred from individual-level posterior parameter estimates using an iterative variational Bayes algorithm that alternates between estimating the population parameters and the individual subject parameters. Over algorithm iterations, individual-level priors are shrunk toward the inferred parent population distribution, as in standard multilevel regression. Furthermore, to reduce the possibility that individual differences in voxelwise estimates from model-based fMRI analyses reflected differences in the scaling of SCEPTIC parameters, we refit the SCEPTIC model to participant data at the group mean parameter values. This approach supports comparisons of regression coefficients across subjects and reduces the confounding of brain-behavior analyses by the individual fits of the computational model to a participant's behavior. We note, however, that our results were qualitatively the same when model

parameters were free to vary across people (additional details available from the corresponding author upon request).

fMRI analyses

Voxelwise fMRI general linear model analyses

Voxelwise general linear model (GLM) analyses of fMRI data were performed using FSL version 6.0.4⁷⁹. Single-run analyses were conducted using FSL FILM, which implements an enhanced version of the GLM that corrects for temporal autocorrelation by prewhitening voxelwise time series and regressors in the design matrix⁸⁵. For each design effect, we convolved a duration-modulated unitheight boxcar regressor with a canonical double-gamma hemodynamic response function (HRF) to yield the model-predicted BOLD response. All models included convolved regressors for the clock and feedback phases of the task.

Moreover, GLM analyses included parametric regressors derived from SCEPTIC. For each whole-brain analysis, we added a single model-based regressor from SCEPTIC alongside the clock and feedback regressors. Results were qualitatively unchanged, however, when all SCEPTIC signals were included as simultaneous predictors, given the relatively low correlation among these signals. For each model-based regressor, the SCEPTIC-derived signal was mean-centered prior to convolution with the HRF. The reward prediction error and entropy change signals were aligned with the feedback, whereas entropy was aligned with the clock (decision) phase. Furthermore, for regressors aligned with the clock phase, which varied in duration, we sought to unconfound the height of the predicted BOLD response due to decision time from the parametric influence of the SCEPTIC signal. Toward this end, for each trial, we convolved a duration-modulated boxcar with the HRF, renormalized the peak to unity, multiplied the regressor by the SCEPTIC signal on that trial, then summed across trials to derive a single model-based regressor (cf. processing time versus intensity of activation in⁹⁵).

Parameter estimates from each run were combined using a weighted fixed effects model in FEAT that propagated error variances from the individual runs. The contrasts from the second-level analyses were then analyzed at the group level using a mixed effects approach implemented in FSL FLAME. Specifically, we used the FLAME 1+2 approach with automatic outlier deweighting⁹⁶, which implements Bayesian mixed effects estimation of the group parameter estimates including full Markov Chain Monte Carlo-based estimation for near-threshold voxels⁹⁷. To identify statistical parametric maps that best represented the average response, all group analyses included age and sex as covariates of no interest (esp. given the developmental sample).

Although our analyses focus primarily on the dorsal attention network (DAN) as the a priori network of interest, we nevertheless conducted whole-brain corrections to the voxelwise GLM statistics to examine the pattern of activity for the signals of interest. Specifically, to correct for familywise error at the whole-brain level, we applied the probabilistic threshold-free cluster enhancement methods p^{TFCE_i} 9⁸, thresholding whole-brain maps at *FWE p* < .05 (e.g., Figure 2B). This algorithm provides strict control over familywise error and boosts sensitivity to clusters of activated voxels.

Brain-behavior fMRI analyses using regression coefficients from model-based fMRI GLM analyses

To relate individual differences in entropy- and entropy change-related BOLD modulation to behavior on the clock task, we extracted subject-level parameter estimates for these GLM contrasts from each of the 47 DAN parcels defined by the Schaefer cortical parcellation ^{see Table s1;,34}. These parameter estimates (aka "betas") served as individual difference measures of sensitivity to signals from SCEPTIC — particularly entropy and entropy change — across regions in the DAN.

We then entered DAN betas for SCEPTIC entropy change as a cross-level moderator of trial-level effects in multilevel models of behavior. Specifically, the dependent variable was trial-wise RT (choice) with behavioral variables as predictors. All models included the trial (inverse-transformed) and previous reward as covariates. Models also included the influence of previous choice (RT_{t-1}) on current choice (RT_t), or RT autocorrelation. A weaker autocorrelation indicates larger exploratory RT swings, and variables that decrease autocorrelation are considered to increase exploration. Most models included DAN entropy change betas as cross-level moderators of the RT_{t-1} effect as a test of how sensitivity to updates in the number of good options modulated exploration on the task. Likewise, most multilevel models also included the trial-varying location of the best option, RT_{Vmax} . The two-way interaction of RT_{Vmax} and DAN entropy change betas tests whether sensitivity to entropy change enhances or diminishes exploitation of the best option.

Because our behavioral observations had a clustered structure (i.e., trials nested within subjects), we used multilevel regression models, which were estimated using restricted maximum likelihood in the *lme4* package⁹⁹ in *R* 4.2.0¹⁰⁰. Estimated *p*-values for predictors in the model were computed using Wald chi-square tests and degrees of freedom were based on the Kenward-Roger approximation. For trial-level analyses, subject and run were treated as random and random intercepts were included for these factors. Additionally, as noted in Results, we included random slopes of key terms such as RT_{Vmax} and RT_{t-1} to ensure the robustness of DAN modulation of exploitation and exploration¹⁰¹.

Within-trial mixed-effects survival analyses of behavior with time-varying value estimates

To examine the sensitivity of choices to within-trial time-varying value, we performed survival analyses predicting the temporal occurrence of response. These mixed-effects Cox models (R coxme package)¹⁰² estimated response hazard as a function of model-predicted expected value and their interaction with session-level DAN responses. This survival analysis does not assume that one precommits to a given response time, instead modeling the within-trial response hazard function in real, continuous time, accounts for censoring and allows for a completely general baseline hazard function¹⁰³. The survival approach accounts for censoring of later within-trial time points by early responses. Most importantly, it allows for a completely general baseline hazard function that can vary randomly across participants. We thus avoid assumptions about the statistical distribution of response times and account for trial-invariant influences such as urgency, processing speed constraints or opportunity cost. We also modeled only the 1000 – 3500 ms interval, excluding early response times that may be shorter than the deliberation and motor planning period and the end of the interval which one may avoid in order to not miss responding on a trial. We included learned value from the information-compressing model as a time-varying covariate, sampled every 100 ms. To account for between-persons heterogeneity, person-specific intercept was included as a random effects; sensitivity analyses also included the random slope of the predictor of interest (value).

Analyses of within-trial peri-feedback BOLD responses using voxelwise deconvolution

Although betas from fMRI GLMs provide a useful window into how decision signals from SCEPTIC relate to behavior at the level of an entire session, the GLM approach makes a number of assumptions: a) that one correctly specifies when in time a signal derived from a computational model modulates neural activity, b) that there is a linear relationship between the model signal and BOLD activity, and c) that a canonical HRF describes the BOLD activity corresponding to a given model-based signal. Furthermore, a conventional model-based fMRI GLM does not allow one to interrogate whether the representation of a given cognitive process varies in time over the course of a trial. For these reasons, we conducted additional analyses that could provide a detailed view of how DAN activity changes following feedback on each trial of the clock task. These analyses also attempted to overcome statistical and conceptual limitations of the GLM and to provide an index of within-trial neural activity that was independent of our computational model. That is, in these analyses, within-trial BOLD activity is the dependent variable and parameters from the SCEPTIC model are predictors.

We first applied a leading hemodynamic deconvolution algorithm to estimate neural activity from BOLD data¹⁰⁴. This algorithm has performed better than alternatives in simulated and real fMRI data, and it is reasonably robust to variations in the timing of neural events and the sampling frequency of the scan¹⁰⁵. We deconvolved the voxelwise BOLD activity for all subjects, averaged the deconvolved time series within each of the 47 DAN parcels (Table s1), and retained these as a regions x time matrix for each run of fMRI data.

Then, to estimate neural activity for each trial in the experiment, we extracted the deconvolved signal surrounding feedback onset (-4 to +4 seconds), censoring timepoints that intersected the previous or next trials. Finally, to ensure that discrete-time models of neural activity could be easily applied, we resampled deconvolved neural activity onto an evenly spaced grid aligned to the feedback onset using linear interpolation. The sampling frequency of the feedback-aligned deconvolved signals was matched to the TR of the fMRI scan (1s for the original sample and 0.6s for the replication sample). Thus, this interpolation was a form of resampling, but did not upsample or downsample the data in the time domain.

For each subject, this yielded a 400 trial x 9 time point (-4—4s for the main sample) x 47 region matrix. We then concatenated these matrices across participants for group analysis. Our primary analyses focused on the four parcels of the DAN visuomotor gradient (MT+, Caudal PPC, Rostral PPC, Premotor), rather than analyzing each region separately. Within each time x parcel combination, we regressed trial-wise neural activity on key decision variables in a multilevel regression framework implemented in *lme4*⁹⁹ in *R*, allowing for crossed random intercepts of subject and side (right/left). Within this framework, the regression coefficients provide an estimate of when and in what region key signals such as entropy change are associated with feedback-related changes in neural activity. Critically, however, given the temporal smoothness of BOLD data, the deconvolved signals remain highly autocorrelated and we are cautious about overinterpreting the temporal precision of these analyses. Moreover, this temporal (and potentially spatial) association results in non-independent statistical tests across the set of space x time models. To adjust for multiple comparisons in non-independent models, we applied the Benjamini–Yekutieli correction across terms of interest in these

models to maintain a false discovery rate of $.05^{106}$.

Another advantage of this analytic approach is that alternative models of representation and behavior can be compared in terms of their alignment to neural activity in fMRI. More specifically, each multilevel model across the space x time set of models yields global fit measures such as the Akaike Information Criterion (AIC), which can be used to compare the relative fit of cognitive signals (e.g., entropy change) to event-aligned BOLD data. Here, we used a global model selection approach¹⁰⁷ based on the AIC to compare the fit of information-compressing, traditional RL, and working memory accounts of the clock task to activity in the DAN (Figure 3).

MEG Analyses

Multi-level analyses of time-frequency domain MEG data

The goal of these analyses was to estimate how reinforcement modulated oscillatory power at each within-trial timepoint and each frequency. To estimate this effect accurately and robustly across sensors and individuals, we used high-performance parallel computing to fit one multi-level linear model for each point in this time-frequency space, combining data from all trials, individuals, and sensors. Predictors included the SCEPTIC model-derived entropy change signal and behavioral confounds: current and previous response times, reward/omission, trial and, in sensitivity analyses, the KL distance between the last and three preceding response times to account for stochastic choice histories. Subject and sensor were treated as crossed random effects, with sensor-specific random intercepts and random slopes of the behavioral variable of interest and subject-specific random intercepts and, where indicated, random slopes of the variable of interest. Since our contrasts were between trials, the intercept accounted for marginal oscillatory power at a given time-frequency point, and correction for baseline was not necessary. Models were estimated using restricted maximum likelihood in the *lme4* package⁹⁹ in R 4.2.0¹⁰⁰. Estimated *p*-values for predictors in the model were computed using Wald chi-square tests and degrees of freedom were based on the Kenward-Roger approximation. To examine the anatomical distribution of effects, after obtaining estimates for each subject and sensor within this overall model, we projected them into the sensor space (Figure 5C) and source space (Figure 5D) as follows.

Source location was performed using the linearly constrained minimum variance (LCMV) Beamformer procedure¹⁰⁸. We used Freesurfer's "fsaverage" template source space and sensor-to-template registration provided by the MNE software¹⁰⁹. The forward model was calculated using the single-layer boundary element, for a total of 20,484 potential source locations placed with 5 mm spacing on the fsaverage surface. A spatial filter was then constructed using a unit-gain LCMV Beamformer¹⁰⁸, with covariances estimated using the 1-second window from the peristimulus interval and 1-second window after the feedback presentation, across all trials and all subjects. We applied the filter to project sensor-level group statistics to the source space. Source estimates were thresholded from 20th to 95th percentiles.

Our analyses of the relationship between subject-level oscillatory responses and behavioral exploration/exploitation employed multi-level survival models identical to those described above

(fMRI Analyses, Within-trial mixed-effects survival analyses of behavior with time-varying value estimates).

Author contributions

Conceptualization: AYD, MNH. Data curation: KH, MNH. Formal analysis: AYD, KH, MNH. Funding acquisition: AYD, BL, MNH. Investigation: KH, MNH. Methodology: AYD, KH, MNH. Project administration: AYD, BL, MNH. Resources: AYD, BL, MNH. Software: MNH. Visualization: AYD, KH, MNH. Writing – original draft: AYD, MNH. Writing – review & editing: AYD, BL, KH, MNH.

Acknowledgments

This work was funded by Ko1 MHo97091, Ro1 MHo67924, and Ro1MH10095 from the National Institute of Mental Health.

The authors thank Rajpreet Chahal, Mandy Collier, Tanya Shah, Shreya Sheth, Laura Taglioni (data collection), Jiazhou Chen, Bea Langer, and Angela Ianni (data processing and analyses). The authors also thank Carl Olson for helpful comments on an earlier draft of the manuscript.

References

- 1. Blanchard, T.C., and Gershman, S.J. (2018). Pure correlates of exploration and exploitation in the human brain. Cogn. Affect. Behav. Neurosci. *18*, 117–126. 10.3758/s13415-017-0556-2.
- Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A.R., and Khamassi, M. (2018). Dopamine regulates the exploration-exploitation trade-off in rats (Neuroscience) 10.1101/482802.
- 3. Costa, V.D., Mitz, A.R., and Averbeck, B.B. (2019). Subcortical Substrates of Explore-Exploit Decisions in Primates. Neuron *103*, 533-545.e5. 10.1016/j.neuron.2019.05.017.
- 4. Frank, M.J., Doll, B.B., Oas-Terpstra, J., and Moreno, F. (2009). The neurogenetics of exploration and exploitation: Prefrontal and striatal dopaminergic components. Nat. Neurosci. *12*, 1062–1068.
- 5. Dombrovski, A.Y., Luna, B., and Hallquist, M.N. (2020). Differential reinforcement encoding along the hippocampal long axis helps resolve the explore—exploit dilemma. Nat. Commun. 11, 5407. 10.1038/s41467-020-18864-0.
- Leonard, T.K., Mikkila, J.M., Eskandar, E.N., Gerrard, J.L., Kaping, D., Patel, S.R., Womelsdorf, T., and Hoffman, K.L. (2015). Sharp Wave Ripples during Visual Exploration in the Primate Hippocampus. J. Neurosci. 35, 14771–14782. 10.1523/JNEUROSCI.0864-15.2015.
- Mumby, D.G., Gaskin, S., Glenn, M.J., Schramek, T.E., and Lehmann, H. (2002). Hippocampal Damage and Exploratory Preferences in Rats: Memory for Objects, Places, and Contexts. Learn. Mem. 9, 49–57. 10.1101/lm.41302.
- 8. Sussman, R.W., Tab Rasmussen, D., and Raven, P.H. (2013). Rethinking Primate Origins Again. Am. J. Primatol. *75*, 95–106. 10.1002/ajp.22096.

- 9. Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. Philos. Trans. R. Soc. B Biol. Sci. *362*, *1585–1599*. *10.1098*/rstb.2007.2054.
- 10. Freedman, D.J., and Ibos, G. (2018). An Integrative Framework for Sensory, Motor, and Cognitive Functions of the Posterior Parietal Cortex. Neuron *97*, 1219–1234. 10.1016/j.neuron.2018.01.044.
- 11. Sereno, M.I., and Huang, R.-S. (2014). Multisensory maps in parietal cortex. Curr. Opin. Neurobiol. 24, 39–46. 10.1016/j.conb.2013.08.014.
- 12. Collins, A.G.E. (2018). The Tortoise and the Hare: Interactions between Reinforcement Learning and Working Memory. J. Cogn. Neurosci. *30*, 1422–1432. 10.1162/jocn_a_01238.
- 13. Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J.P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. Neuron *66*, 585–595. 10.1016/j.neuron.2010.04.016.
- Leong, Y.C., Radulescu, A., Daniel, R., DeWoskin, V., and Niv, Y. (2017). Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. Neuron *93*, 451–463. 10.1016/j.neuron.2016.12.040.
- 15. Niv, Y., Daniel, R., Geana, A., Gershman, S.J., Leong, Y.C., Radulescu, A., and Wilson, R.C. (2015). Reinforcement Learning in Multidimensional Environments Relies on Attention Mechanisms. J. Neurosci. 35, 8145–8157. 10.1523/JNEUROSCI.2978-14.2015.
- Anderson, B.A. (2017). Reward processing in the value-driven attention network: reward signals tracking cue identity and location. Soc. Cogn. Affect. Neurosci. 12, 461–467.
 10.1093/scan/nsw141.
- 17. Foster, J.J., Bsales, E.M., Jaffe, R.J., and Awh, E. (2017). Alpha-Band Activity Reveals Spontaneous Representations of Spatial Position in Visual Working Memory. Curr. Biol. 27, 3216-3223.e6. 10.1016/j.cub.2017.09.031.
- 18. Gelastopoulos, A., Whittington, M.A., and Kopell, N.J. (2019). Parietal low beta rhythm provides a dynamical substrate for a working memory buffer. Proc. Natl. Acad. Sci. 116, 16613–16620. 10.1073/pnas.1902305116.
- 19. Peck, C.J., Jangraw, D.C., Suzuki, M., Efem, R., and Gottlieb, J. (2009). Reward Modulates Attention Independently of Action Value in Posterior Parietal Cortex. J. Neurosci. 29, 11182– 11191. 10.1523/JNEUROSCI.1929-09.2009.
- 20. Hallquist, M.N., and Dombrovski, A.Y. (2019). Selective maintenance of value information helps resolve the exploration/exploitation dilemma. Cognition *183*, 226–243. 10.1016/j.cognition.2018.11.004.
- 21. Ahmed, Z., Roux, N.L., Norouzi, M., and Schuurmans, D. (2019). Understanding the Impact of Entropy on Policy Optimization. In Proceedings of the 36th International Conference on Machine Learning (PMLR), pp. 151–160.
- 22. Sutton, R.S., and Barto, A.G. (2018). Reinforcement Learning, second edition: An Introduction (MIT Press).
- 23. Brown, V.M., Hallquist, M.N., Frank, M.J., and Dombrovski, A.Y. (2022). Humans adaptively resolve the explore-exploit dilemma under cognitive constraints: Evidence from a multi-armed bandit task. Cognition *229*, 105233. 10.1016/j.cognition.2022.105233.

- 24. Woodford, M. (2009). Information-constrained state-dependent pricing. J. Monet. Econ. *56*, S100–S124. 10.1016/j.jmoneco.2009.06.014.
- 25. Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238. 10.1038/22268.
- 26. Leathers, M.L., and Olson, C.R. (2012). In monkeys making value-based decisions, LIP neurons encode cue salience and not action value. Science 338, 132–135. 10.1126/science.1226405.
- 27. Fischer, P., Tan, H., Pogosyan, A., and Brown, P. (2016). High post-movement parietal low-beta power during rhythmic tapping facilitates performance in a stop task. Eur. J. Neurosci. 44, 2202–2213. 10.1111/ejn.13328.
- 28. Savoie, F.-A., Thénault, F., Whittingstall, K., and Bernier, P.-M. (2018). Visuomotor Prediction Errors Modulate EEG Activity Over Parietal Cortex. Sci. Rep. *8*, 12513. 10.1038/s41598-018-30609-0.
- 29. Sutterer, D.W., Foster, J.J., Serences, J.T., Vogel, E.K., and Awh, E. (2019). Alpha-band oscillations track the retrieval of precise spatial representations from long-term memory. J. Neurophysiol. *122*, 539–551. 10.1152/jn.00268.2019.
- Fiebelkorn, I.C., and Kastner, S. (2021). Spike Timing in the Attention Network Predicts Behavioral Outcome Prior to Target Selection. Neuron *109*, 177-188.e4.
 10.1016/j.neuron.2020.09.039.
- 31. Ludvig, E.A., Sutton, R.S., and Kehoe, E.J. (2008). Stimulus Representation and the Timing of Reward-Prediction Errors in Models of the Dopamine System. Neural Comput. *20*, 3034–3054. 10.1162/neco.2008.11-07-654.
- 32. Ludvig, E.A., Sutton, R.S., and Kehoe, E.J. (2012). Evaluating the TD model of classical conditioning. Learn. Behav. *40*, 305–319. 10.3758/s13420-012-0082-6.
- 33. Thomas Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. *106*, 1125– 1165. 10.1152/jn.00338.2011.
- 34. Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., and Yeo, B.T.T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb. Cortex *28*, 3095–3114. 10.1093/cercor/bhx179.
- 35. Beckmann, C.F., DeLuca, M., Devlin, J.T., and Smith, S.M. (2005). Investigations into restingstate connectivity using independent component analysis. Philos. Trans. R. Soc. B Biol. Sci. *360*, 1001–1013. 10.1098/rstb.2005.1634.
- 36. Bray, S., Arnold, A.E.G.F., Levy, R.M., and Iaria, G. (2015). Spatial and temporal functional connectivity changes between resting and attentive states: Connectivity Changes Between Rest and Attention. Hum. Brain Mapp. *36*, 549–565. 10.1002/hbm.22646.
- 37. Astafiev, S.V., Shulman, G.L., Stanley, C.M., Snyder, A.Z., Van Essen, D.C., and Corbetta, M. (2003). Functional Organization of Human Intraparietal and Frontal Cortex for Attending, Looking, and Pointing. J. Neurosci. 23, 4689–4699. 10.1523/JNEUROSCI.23-11-04689.2003.

- 38. Scheperjans, F., Eickhoff, S.B., Hömke, L., Mohlberg, H., Hermann, K., Amunts, K., and Zilles, K. (2008). Probabilistic Maps, Morphometry, and Variability of Cytoarchitectonic Areas in the Human Superior Parietal Cortex. Cereb. Cortex *18*, 2141–2157. 10.1093/cercor/bhm241.
- 39. Sakreida, K., Effnert, I., Thill, S., Menz, M.M., Jirak, D., Eickhoff, C.R., Ziemke, T., Eickhoff, S.B., Borghi, A.M., and Binkofski, F. (2016). Affordance processing in segregated parieto-frontal dorsal stream sub-pathways. Neurosci. Biobehav. Rev. 69, 89–112. 10.1016/j.neubiorev.2016.07.032.
- Davare, M., Rothwell, J.C., and Lemon, R.N. (2010). Causal Connectivity between the Human Anterior Intraparietal Area and Premotor Cortex during Grasp. Curr. Biol. 20, 176–181. 10.1016/j.cub.2009.11.063.
- Field, D.T., Biagi, N., and Inman, L.A. (2020). The role of the ventral intraparietal area (VIP/pVIP) in the perception of object-motion and self-motion. NeuroImage 213, 116679.
 10.1016/j.neuroimage.2020.116679.
- 42. Verhagen, L., Dijkerman, H.C., Grol, M.J., and Toni, I. (2008). Perceptuo-Motor Interactions during Prehension Movements. J. Neurosci. *28*, 4726–4735. 10.1523/JNEUROSCI.0057-08.2008.
- 43. Richter, M., Amunts, K., Mohlberg, H., Bludau, S., Eickhoff, S.B., Zilles, K., and Caspers, S. (2019). Cytoarchitectonic segregation of human posterior intraparietal and adjacent parieto-occipital sulcus and its relation to visuomotor and cognitive functions. Cereb. Cortex 29, 1305–1327. 10.1093/cercor/bhy245.
- 44. Rizzolatti, G., and Matelli, M. (2003). Two different streams form the dorsal visual system: anatomy and functions. Exp. Brain Res. *153*, 146–157. 10.1007/s00221-003-1588-0.
- 45. Failing, M., and Theeuwes, J. (2018). Selection history: How reward modulates selectivity of visual attention. Psychon. Bull. Rev. 25, 514–538. 10.3758/s13423-017-1380-y.
- 46. Cavanagh, J.F., Figueroa, C.M., Cohen, M.X., and Frank, M.J. (2012). Frontal Theta Reflects Uncertainty and Unexpectedness during Exploration and Exploitation. Cereb. Cortex 22, 2575– 2586. 10.1093/cercor/bhr332.
- 47. Bolger, N., Zee, K.S., Rossignac-Milon, M., and Hassin, R.R. (2019). Causal processes in psychology are heterogeneous. J. Exp. Psychol. Gen. *148*, 601–618. 10.1037/xge0000558.
- 48. Ahmad, S. (1991). VISIT: An efficient computational model of human visual attention.
- 49. Bisley, J.W., and Mirpour, K. (2019). The neural instantiation of a priority map. Curr. Opin. Psychol. 29, 108–112. 10.1016/j.copsyc.2019.01.002.
- 50. Colby, C.L., and Goldberg, M.E. (1999). Space and Attention in Parietal Cortex. Annu. Rev. Neurosci. 22, 319–349. 10.1146/annurev.neuro.22.1.319.
- 51. Serences, J.T., and Yantis, S. (2004). Attentional Priority Maps in Human Cortex: (537052012-668). 10.1037/e537052012-668.
- 52. Snyder, L.H., Batista, A.P., and Andersen, R.A. (2000). Intention-related activity in the posterior parietal cortex: a review. Vision Res. *40*, 1433–1441. 10.1016/S0042-6989(00)00052-3.
- 53. Polanía, R., Moisa, M., Opitz, A., Grueschow, M., and Ruff, C.C. (2015). The precision of valuebased choices depends causally on fronto-parietal phase coupling. Nat. Commun. 6, 8090. 10.1038/ncomms9090.

- 54. Horan, M., Daddaoua, N., and Gottlieb, J. (2019). Parietal neurons encode information sampling based on decision uncertainty. Nat. Neurosci. 22, 1327–1335. 10.1038/s41593-019-0440-1.
- 55. Li, Y., Daddaoua, N., Horan, M., Foley, N.C., and Gottlieb, J. (2022). Uncertainty modulates visual maps during noninstrumental information demand. Nat. Commun. *13*, 5911. 10.1038/s41467-022-335⁸5-2.
- 56. Gibson, J.J. (1977). The theory of affordances. Hilldale USA 1, 67–82.
- 57. Calton, J.L., Dickinson, A.R., and Snyder, L.H. (2002). Non-spatial, motor-specific activation in posterior parietal cortex. Nat. Neurosci. *5*, 580–588. 10.1038/nno602-862.
- 58. Tseng, S.-Y., Chettih, S.N., Arlt, C., Barroso-Luque, R., and Harvey, C.D. (2022). Shared and specialized coding across posterior cortical areas for dynamic navigation decisions. Neuron *110*, 2484-2502.e16. 10.1016/j.neuron.2022.05.012.
- 59. Zhang, C.Y., Aflalo, T., Revechkis, B., Rosario, E.R., Ouellette, D., Pouratian, N., and Andersen, R.A. (2017). Partially Mixed Selectivity in Human Posterior Parietal Association Cortex. Neuron *95*, 697-708.e4. 10.1016/j.neuron.2017.06.040.
- 60. Cicmil, N., Cumming, B.G., Parker, A.J., and Krug, K. (2015). Reward modulates the effect of visual cortical microstimulation on perceptual decisions. eLife 4, e07832. 10.7554/eLife.07832.
- 61. Caligiore, D., Borghi, A.M., Parisi, D., and Baldassarre, G. (2010). TROPICALS: A computational embodied neuroscience model of compatibility effects. Psychol. Rev. 117, 1188–1228. 10.1037/a0020887.
- 62. Isaacson, J.S., and Scanziani, M. (2011). How Inhibition Shapes Cortical Activity. Neuron 72, 231– 243. 10.1016/j.neuron.2011.09.027.
- 63. Cisek, P., and Kalaska, J.F. (2010). Neural Mechanisms for Interacting with a World Full of Action Choices. Annu. Rev. Neurosci. *33*, 269–298. 10.1146/annurev.neuro.051508.135409.
- 64. Pastor-Bernier, A., and Cisek, P. (2011). Neural Correlates of Biased Competition in Premotor Cortex. J. Neurosci. 31, 7083–7088. 10.1523/JNEUROSCI.5681-10.2011.
- 65. Thill, S., Caligiore, D., Borghi, A.M., Ziemke, T., and Baldassarre, G. (2013). Theories and computational models of affordance and mirror systems: An integrative review. Neurosci. Biobehav. Rev. 37, 491–521. 10.1016/j.neubiorev.2013.01.012.
- 66. Todd, J.J., and Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. Nature *428*, 751–754. 10.1038/nature02466.
- 67. Collins, A.G.E., and Frank, M.J. (2018). Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc. Natl. Acad. Sci. 115, 2502–2507. 10.1073/pnas.1720963115.
- 68. Cisek, P. (2019). Resynthesizing behavior through phylogenetic refinement. Atten. Percept. Psychophys. *81*, 2265–2287. 10.3758/s13414-019-01760-1.
- 69. Adesnik, H., and Scanziani, M. (2010). Lateral competition for cortical space by layer-specific horizontal circuits. Nature *464*, 1155–1160. 10.1038/nature08935.
- 70. Mysore, S.P., and Kothari, N.B. (2020). Mechanisms of competitive selection: A canonical neural circuit framework. eLife *9*, e51473. 10.7554/eLife.51473.

- 71. Magri, C., Schridde, U., Murayama, Y., Panzeri, S., and Logothetis, N.K. (2012). The Amplitude and Timing of the BOLD Signal Reflects the Relationship between Local Field Potential Power at Different Frequencies. J. Neurosci. 32, 1395–1407. 10.1523/JNEUROSCI.3985-11.2012.
- 72. Scheeringa, R., Fries, P., Petersson, K.-M., Oostenveld, R., Grothe, I., Norris, D.G., Hagoort, P., and Bastiaansen, M.C.M. (2011). Neuronal Dynamics Underlying High- and Low-Frequency EEG Oscillations Contribute Independently to the Human BOLD Signal. Neuron 69, 572–583. 10.1016/j.neuron.2010.11.044.
- 73. Zumer, J.M., Brookes, M.J., Stevenson, C.M., Francis, S.T., and Morris, P.G. (2010). Relating BOLD fMRI and neural oscillations through convolution and optimal linear weighting. NeuroImage 49, 1479–1489. 10.1016/j.neuroimage.2009.09.020.
- 74. Essex, B.G., Clinton, S.A., Wonderley, L.R., and Zald, D.H. (2012). The Impact of the Posterior Parietal and Dorsolateral Prefrontal Cortices on the Optimization of Long-Term versus Immediate Value. J. Neurosci. 32, 15403–15413. 10.1523/JNEUROSCI.6106-11.2012.
- 75. Tottenham, N., Tanaka, J.W., Leon, A.C., McCarry, T., Nurse, M., Hare, T.A., Marcus, D.J., Westerlund, A., Casey, B., and Nelson, C. (2009). The NimStim set of facial expressions: judgments from untrained research participants. Psychiatry Res. 168, 242–249.
- 76. Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage 47, S102. 10.1016/S1053-8119(09)70884-5.
- 77. Millman, K.J., and Brett, M. (2007). Analysis of functional magnetic resonance imaging in Python. Comput. Sci. Eng. 9, 52–55.
- 78. Cox, R.W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. Comput. Biomed. Res. 29, 162–173. 10.1006/cbmr.1996.0014.
- 79. Smith, S.M., Jenkinson, M., Woolrich, M.W., Beckmann, C.F., Behrens, T.E.J., Johansen-Berg, H., Bannister, P.R., De Luca, M., Drobnjak, I., Flitney, D.E., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 Suppl 1, S208-219. 10.1016/j.neuroimage.2004.07.051.
- Roche, A. (2011). A Four-Dimensional Registration Algorithm With Application to Joint Correction of Motion and Slice Timing in fMRI. IEEE Trans. Med. Imaging 30, 1546–1554. 10.1109/TMI.2011.2131152.
- 81. Iglesias, J.E., Cheng-Yi Liu, Thompson, P.M., and Zhuowen Tu (2011). Robust Brain Extraction Across Datasets and Comparison With Publicly Available Methods. IEEE Trans. Med. Imaging *30*, 1617–1634. 10.1109/TMI.2011.2138152.
- 82. Greve, D.N., and Fischl, B. (2009). Accurate and robust brain image alignment using boundarybased registration. NeuroImage *48*, 63–72. 10.1016/j.neuroimage.2009.06.060.
- 83. Pruim, R.H.R., Mennes, M., van Rooij, D., Llera, A., Buitelaar, J.K., and Beckmann, C.F. (2015). ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage 112, 267–277. 10.1016/j.neuroimage.2015.02.064.
- 84. Ciric, R., Wolf, D.H., Power, J.D., Roalf, D.R., Baum, G.L., Ruparel, K., Shinohara, R.T., Elliott, M.A., Eickhoff, S.B., Davatzikos, C., et al. (2017). Benchmarking of participant-level confound

regression strategies for the control of motion artifact in studies of functional connectivity. NeuroImage *154*, 174–187. 10.1016/j.neuroimage.2017.03.020.

- 85. Woolrich, M.W., Ripley, B.D., Brady, M., and Smith, S.M. (2001). Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14, 1370–1386. 10.1006/nimg.2001.0931.
- 86. Taulu, S., and Hari, R. (2009). Removal of magnetoencephalographic artifacts with temporal signal-space separation: Demonstration with single-trial auditory-evoked responses. Hum. Brain Mapp. *30*, 1524–1534. 10.1002/hbm.20627.
- 87. Taulu, S., Kajola, M., and Simola, J. (2004). Suppression of Interference and Artifacts by the Signal Space Separation Method. Brain Topogr. *16*, 269–275. 10.1023/B:BRAT.0000032864.93890.f9.
- 88. Nenonen, J., Nurminen, J., Kičić, D., Bikmullina, R., Lioumis, P., Jousmäki, V., Taulu, S., Parkkonen, L., Putaala, M., and Kähkönen, S. (2012). Validation of head movement correction and spatiotemporal signal space separation in magnetoencephalography. Clin. Neurophysiol. 123, 2180–2191. 10.1016/j.clinph.2012.03.080.
- 89. Bell, A.J., and Sejnowski, T.J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput. 7, 1129–1159. 10.1162/neco.1995.7.6.1129.
- 90. Inman, H.F., and Bradley, E.L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. Commun. Stat. Theory Methods *18*, 3851–3874. 10.1080/03610928908830127.
- 91. Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: An introduction (MIT Press).
- 92. Bennett, D., Niv, Y., and Langdon, A. (2021). Value-free reinforcement learning: Policy optimization as a minimal model of operant behavior (PsyArXiv) 10.31234/osf.io/ew58m.
- 93. Sutton, R.S., McAllester, D., Singh, S., and Mansour, Y. (1999). Policy Gradient Methods for Reinforcement Learning with Function Approximation. In Advances in Neural Information Processing Systems (MIT Press).
- 94. Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: A Probabilistic Treatment of Nonlinear Models for Neurobiological and Behavioural Data. PLOS Comput Biol 10, e1003441. 10.1371/journal.pcbi.1003441.
- 95. Poldrack, R.A. (2015). Is "efficiency" a useful concept in cognitive neuroscience? Dev. Cogn. Neurosci. 11, 12–17. 10.1016/j.dcn.2014.06.001.
- 96. Woolrich, M. (2008). Robust group analysis using outlier inference. NeuroImage 41, 286–301. 10.1016/j.neuroimage.2008.02.042.
- 97. Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., Jenkinson, M., and Smith, S.M. (2004). Multilevel linear modelling for FMRI group analysis using Bayesian inference. NeuroImage 21, 1732–1747. 10.1016/j.neuroimage.2003.12.023.
- 98. Spisák, T., Spisák, Z., Zunhammer, M., Bingel, U., Smith, S., Nichols, T., and Kincses, T. (2019). Probabilistic TFCE: A generalized combination of cluster size and voxel intensity to increase statistical power. NeuroImage 185, 12–26. 10.1016/j.neuroimage.2018.09.078.
- 99. Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using Ime4. J. Stat. Softw. *67*, 1–48. 10.18637/jss.vo67.io1.

- 100. R Core Team (2022). R: A language and environment for statistical computing.
- 101. Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing Type I error and power in linear mixed models. J. Mem. Lang. *94*, 305–315. 10.1016/j.jml.2017.01.001.
- 102. Therneau, T.M. (2018). coxme: Mixed Effects Cox Models.
- 103. Singer, J.D., and Willett, J.B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence (Oxford University Press).
- 104. Bush, K., and Cisler, J. (2013). Decoding neural events from fMRI BOLD signal: A comparison of existing approaches and development of a new algorithm. Magn. Reson. Imaging 31, 976–989. 10.1016/j.mri.2013.03.015.
- 105. Bush, K., Cisler, J., Bian, J., Hazaroglu, G., Hazaroglu, O., and Kilts, C. (2015). Improving the precision of fMRI BOLD signal deconvolution with implications for connectivity analysis. Magn. Reson. Imaging 33, 1314–1323. 10.1016/j.mri.2015.07.007.
- 106. Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188. 10.1214/aos/1013699998.
- 107. Burnham, K.P., and Anderson, D.R. (2002). Model selection and multi-model inference: A practical information-theoretic approach 2nd ed. (Springer).
- 108. Van Veen, B.D., Van Drongelen, W., Yuchtman, M., and Suzuki, A. (1997). Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans. Biomed. Eng. 44, 867–880. 10.1109/10.623056.
- 109. Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al. (2013). MEG and EEG data analysis with MNE-Python. Front. Neurosci. 7.