Frontoparietal Hubs Leverage Probabilistic Representations and Integrated Uncertainty to Guide Cognitive Flexibility

Stephanie C. Leach[1], Hannah Hollow[1], Jiefeng Jiang[1,2,¶], and Kai Hwang[1,2, ¶*]

[1] Psychological and Brain Sciences, The University of Iowa, Iowa City, IA.

[2] Cognitive Control Collaborative (CCC).

* Corresponding Author

E-mail: kai-hwang@uiowa.edu

¶ These authors contributed equally to this work.

Abstract

Adaptive behavior requires integrating multiple stochastic sources of information; however, neural and computational mechanisms underlying integration are not fully understood. We introduce a computational neuroimaging framework elucidating how brain systems integrate internally maintained and externally cued information to guide behavior. Our computational model estimates trial-by-trial input beliefs that are integrated into a joint probability distribution, where entropy quantifies overall uncertainty, outputs a probabilistic task belief, and generates continuous measures of prediction errors and updating. Results revealed latent-state beliefs are encoded primarily in default-mode regions, whereas perceptual beliefs in attention and frontoparietal networks; the integrated joint probability and its entropy converged in frontoparietal hub areas, notably middle frontal gyrus and intraparietal sulcus. Entropy adaptively reconfigured connectivity—first strengthening coupling with sensory regions, then task regions, and finally prediction-error circuits. This suggests that frontoparietal hubs read out and resolve distributed uncertainty to flexibly guide behavior, revealing how frontoparietal systems implement cognitive integration.

Introduction

The ability to engage in complex goal-directed behavior is a hallmark of human intelligence. To achieve goals, humans construct rich cognitive representations that capture rule structures and contextual contingencies (1, 2). These cognitive representations transform sensory and mnemonic inputs into actions and decisions to guide goal-directed behavior. Critically, goal directed behavior often involves integrating multiple sources of information, including perceptual inputs and internal beliefs/expectations. For instance, when driving, flashing lights from a distant accident must be combined with internal expectations about traffic to select an optimal alternative route. However, the neural and cognitive processes underlying cognitive integration—combining

abstract beliefs/expectations with perceptual inputs—are not well understood. Cognitive integration is challenging because the sources may be non-observable (e.g., internal thoughts), corrupted by noise, or stochastic in nature. Thus, the mechanism for integrating potentially stochastic information, from observable and non-observable sources, is presently unknown, leaving a critical gap in our understanding of integration mechanisms.

Integration functions are believed to be supported by hub regions in the frontoparietal network (FPN), which exhibit extensive connectivity across multiple systems (3-5). Hubs have been proposed to leverage this connectivity architecture to flexibly receive inputs and transmit outputs among diverse brain networks, thereby supporting integration (6). Studies of functional brain networks often identify regions in the frontoparietal network (FPN) and the default mode network (DMN) as exhibiting strong hub properties (7). However, connectivity patterns alone do not reveal the underlying neurocognitive mechanisms of integration. Specifically, they neither explicitly test the computational processes enabling integration nor clarify the representational structure of encoded cognitive information. Consequently, how hub regions integrate diverse information remains unclear.

To address this issue, we propose a computational framework clarifying the integration of noisy or uncertain information that is either internally maintained or externally cued. This framework specifies: (1) the inputs being integrated, (2) processes transforming these inputs into an integrated representation, (3) the outputs derived from this representation, and (4) updating mechanisms modifying inputs based on the integrated outputs. Critically, incorporating these elements with neuroimaging can precisely probe associated neural processes.

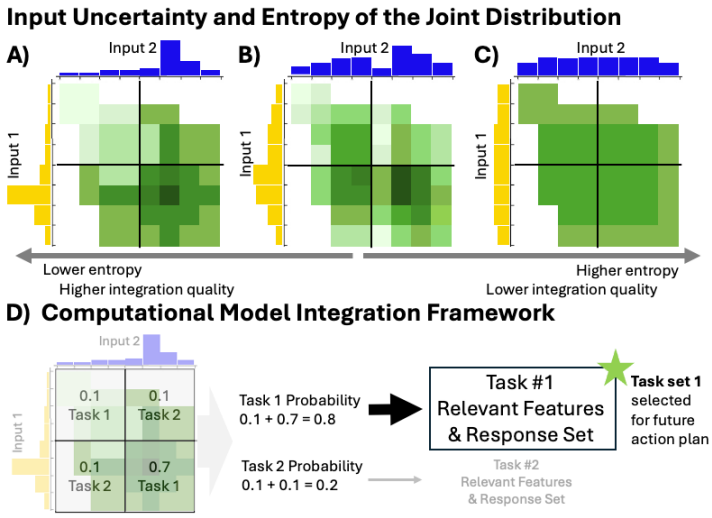

Figure 1. **Integrated Uncertainty**. (**A-C**) Depiction of how integrated uncertainty, or entropy of the joint distribution, fluctuates with different levels of input uncertainty. The shading of the cells in the joint distribution reflect the peak of the integrated distribution. A clear peak, as indicated by a darker shade of green, indicates a low level of integration uncertainty. A lack of a clear peak indicates a high level of integration uncertainty. (**A**) Low levels of input uncertainty lead to higher integration quality and therefore lower levels of integrated uncertainty. (**B**) Moderate levels of input uncertainty lead to moderate levels of integrated uncertainty. (**C**) High levels of input uncertainty lead to a lower integration quality and therefore high levels of integrated uncertainty. (**D**) Depiction of integration framework where joint distribution maps onto a conditional probabilistic distribution to output a single, executable task set.

We developed a computational model based on a Bayesian framework to infer latent variables that represent subjective beliefs. These beliefs are from externally presented and internally maintained information sources (inputs), as well as the output of the integrated representation. This model posits that internal and external sources of information are integrated into a joint probabilistic distribution that maps onto a conditional probabilistic distribution (Fig. 1). The overall uncertainty in integration is measured through the entropy of the joint distribution (Fig. 1A-C). The probability of performing a given action/decision is conditioned on the joint contribution of both external and internal sources (Fig 1D). The latent variables include categorical beliefs of inputs/outputs and their associated uncertainty. Estimating this joint distribution allows for further examination of prediction errors and updating processes. Our framework thus provides trial-by-trial predictors detailing input-output representational structures, integrated representations, uncertainty estimates, and prediction errors driving belief updates. For

this study, we combined these model-derived estimates with advanced fMRI analyses to comprehensively investigate the neurocognitive process of integration.

Results

To address cognitive integration of goal information, we designed a task paradigm in which participants integrated one external and one internal source of noisy information to determine the correct task for each trial (Fig. **2**). The external source was a perceptual array of red and yellow dots (Fig. **2C**), while the internal source was a hidden state that randomly changed every 10–20 trials (Fig. **2A**). Participants performed either the "Face task", a gender judgment, or "Scene task", a place judgement (Fig. **2D**). Participants had to rely on feedback to infer whether the hidden state had changed (Fig. **2E**) because there was no other external indication of state transitions. Cue, probe, and feedback screens were separated by jittered delays (range: 0.5-10.5s, avg.=3.75s; Fig. **2B**).

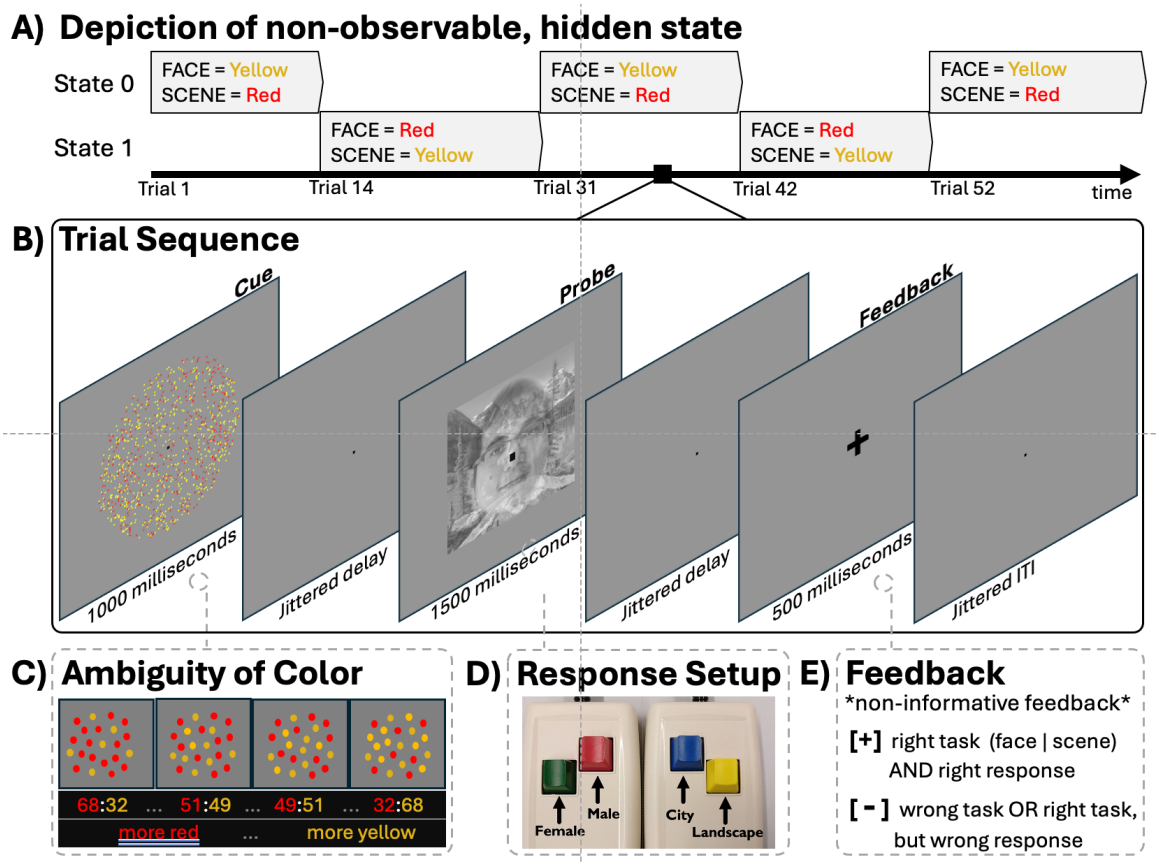

Fig 2. Experimental Design. (**A**) Depiction of non-observable, volatile state that changes every 10-20 trials. **(B)** Trial sequence example: cue screen (1000ms), jittered delay (range: 0.5-10.5 seconds, average=3.75 seconds), probe screen (1500ms), jittered delay (range: 0.5-10.5 seconds, average=3.75 seconds), feedback screen (500ms), and jittered intertrial interval (range: 0.5-10.5 seconds, average=3.75 seconds). Example face stimulus is a photo of the corresponding author of this manuscript. (**C**) Depiction of color ambiguity manipulation. Some trials were less ambiguous with proportions closer to 68:32 while other trials were more ambiguous with proportions closer to 51:49. **(D)** Response setup for fMRI. Note, task was split by hand. (**E**) Possible feedback. Note, feedback was non-informative, meaning this feedback did not clarify if participants were wrong because they picked the wrong gender or scene type or because they did the wrong task.

2.1 Computational model performance

We developed a Bayesian model (Fig. **3A**) to capture the computational processes underlying cognitive integration in this paradigm. This model estimates two probabilistic distributions—the volatile hidden state and the ambiguous perceptual cue, or color ratio— which serve as the inputs that are integrated into a joint probability distribution. This model was applied separately to each subject's behavioral data (see one example subject in Fig. **3B**). By varying the "noisiness" (i.e., ambiguity/volatility) of these inputs across trials, we manipulated integration

uncertainty. The joint probability provided a direct measure of integration while the entropy of the joint distribution offers an indirect measure of overall uncertainty. This integrated uncertainty, or entropy, in turn influenced the model output (i.e., task to be performed). Furthermore, by tracking uncertainty in both inputs and outputs, the model allowed us to investigate how these representations are updated on a trial-by-trial basis using prediction errors and Bayes' rule. In short, the model outlines a clear framework that generates trial-by-trial predictors for input representations, their integrated product (i.e., joint probability distribution), the integrated uncertainty (as indexed by entropy), and the resulting task output representation.

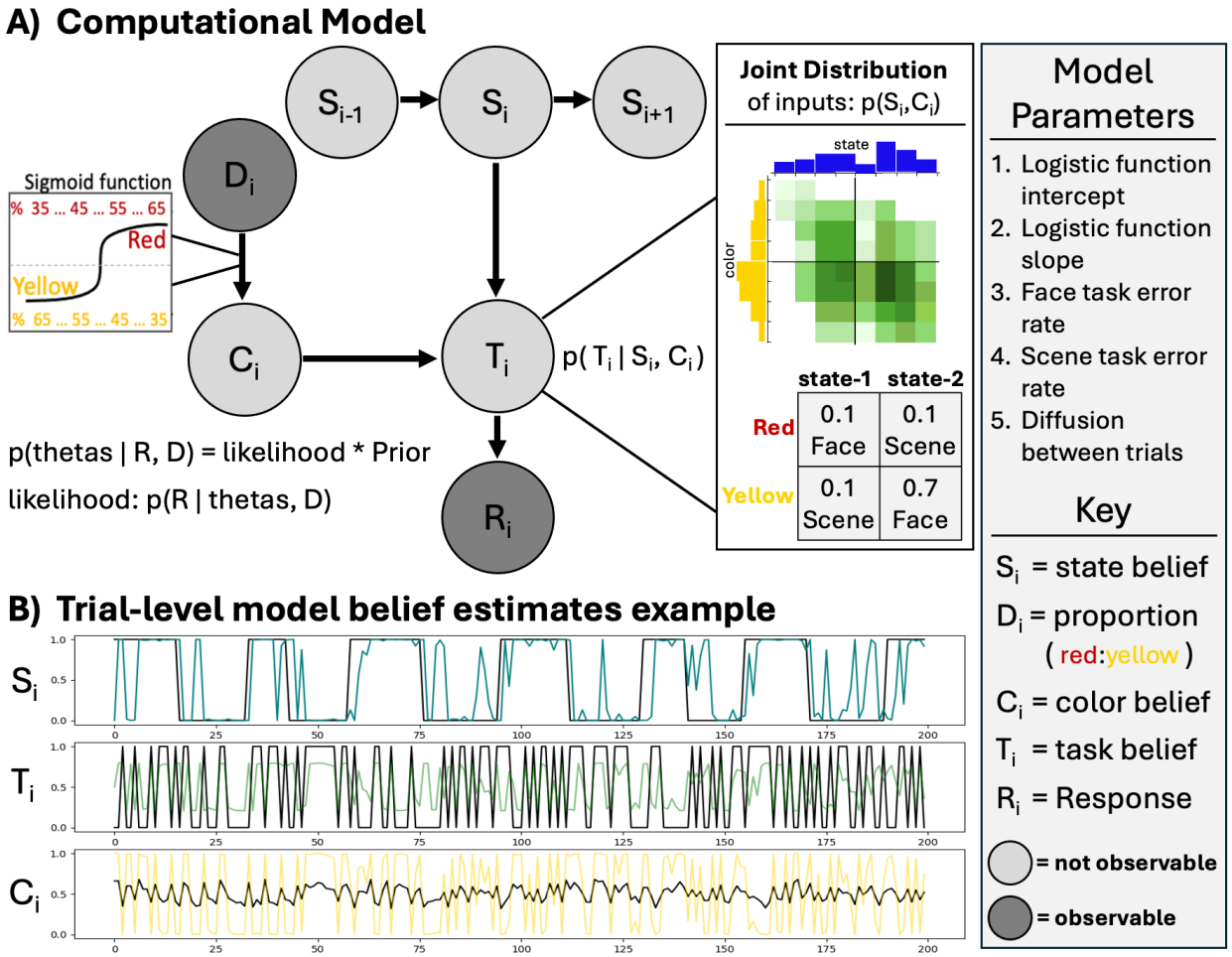

Fig 3. Computational Model. (**A**) Computational (Bayesian generative) model diagram. Observable information (R_i and D_i) used to infer three latent variables (S_i, C_i, and T_i). (**B**) Trial-level model estimated "beliefs" from one example subject. Black lines: true state, task, or color. Blue, green, and yellow lines: state belief, task belief, and color belief, respectively.

The key feature of our proposed model is that both sources of noisy information are encoded probabilistically, enabling their uncertainty to be integrated into a joint representation that predicts task output. To test this probabilistic integration assumption, we compared our model to four alternative models that lacked this feature: (1) dichotomous state and probabilistic color (dSpP), (2) probabilistic state and dichotomous color (pSdP), (3) both state and color dichotomous (pSdP), and (4) a model identical to the. pSdP model but with random guessing under highly ambiguous color conditions (pSjP). This pSjP model was included to capture the possibility that the participant used two different strategies (i.e., random guessing for high color uncertainty and categorical judgments at lower uncertainty).

In these alternative models, representing state or color as dichotomous means beliefs do not include uncertainty but instead a categorical decision. Integration in these models thus occurs as a sequential decision process without forming a full probabilistic joint representation. Consequently, belief updating is also treated as an all-or-none process in which beliefs are either entirely revised or remain unchanged. In contrast, the proposed model accounts for graded uncertainty in both state and color beliefs. If the proposed model, where both sources are probabilistic, outperforms the alternative models, this would support the use of entropy in the joint distribution as a measure of integration in behavioral and neuroimaging analyses.

We fit all models to behavioral data from the 38 usable subjects who completed the fMRI task. Our proposed probabilistic model (pSpP) showed the lowest prediction error rate (23.08%) compared to alternative models (pSdP: 25.03%, dSpP: 33.70%, dSdP: 24.93%, pSjP: 50.04%; Fig. **4A**). Bayesian information criterion further showed strong preference of the proposed integration model (BIC; Fig. **4B**; protected exceedance probability = 0.945, Bayes Omnibus risk < 0.001). Notably, the alternative model assuming a mixture of random guessing and dichotomous color (pSjP) performed the worst, while the other two models assuming dichotomous color (pSdP and dSdP) showed the closest performance to the proposed model. This likely reflects that color

perception is nearly, but not completely, categorical, as indicated by the steep slopes of the sigmoid function used to model color. Overall, these results provide empirical support for the key assumptions of our integrative model.

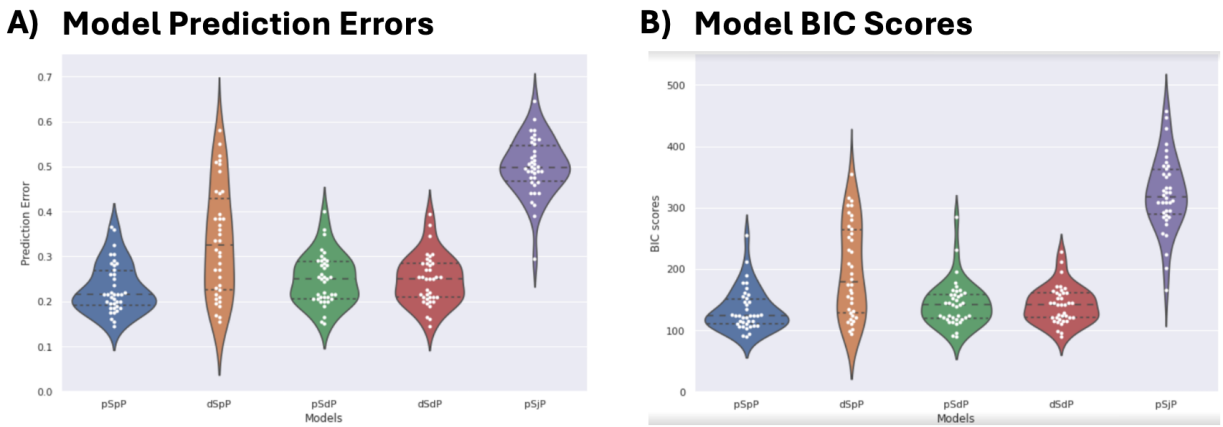

Fig 4. Integration Uncertainty and RTs. (**A**) Prediction error (percent of trials) of each computational model. Each dot represents an individual subject's prediction accuracy. (**B**) BIC scores for each computational model. Each dot represents an individual subject's BIC score.

To further test the behavioral relevance of our model, we performed additional analyses on reaction times (RTs). If entropy of the joint distribution indexes integrated uncertainty, greater entropy should predict slower RTs. To test this, we performed a mixed-effects regression with RT as the dependent variable, entropy as a fixed effect, and participants as random intercepts. Higher entropy significantly predicted slower RTs ($\beta = 0.028$, $t(7572)=3.064$, $p=.002$, Cohen's $d=0.50$). Since task belief derives from this joint distribution, greater task uncertainty should similarly slow RT. A second mixed-effects regression confirmed task uncertainty significantly predicted slower RTs ($\beta = 0.038$, $t(7547)=2.257$, $p=.024$, Cohen's $d=0.37$). Together, these findings support that integrated uncertainty (entropy) and task uncertainty significantly influence response times, thus supporting the behavioral relevance of our computational model.

2.2 Neural substrates of integration

Our model-driven approach enabled us to investigate the neural substrates associated with each component of integration process—inputs, joint representation, output, prediction errors, and updating. Our framework addresses three critical aspects: (1) how distinct information sources are integrated into a unified joint representation, (2) how integration uncertainty, quantified as entropy, biases the resulting task output, and (3) how trial-level feedback drives the updating of input representations. Guided by our model, we ran a series of whole-brain multivariate fMRI analyses to investigate every stage of integration. First, we decoded trial-by-trial beliefs for state, color, and the output task beliefs to identify the neural representations of input and output information. Next, we used model uncertainty estimates to identify regions where activity scaled with input- or output-level uncertainty. We then decoded the joint state–color probability and added its entropy as a parametric regressor to reveal neural representations of the integrated product. To investigate updating processes, we identified structures whose activity varied with state-level prediction errors and belief changes. Finally, we examined how entropy modulated functional connectivity across cue, probe, and feedback epochs. Together, these analyses aimed to reveal the complete cognitive integration process.

2.2.1 Inputs for integration

The inputs for integration were an (1) externally presented color cue and (2) internally maintained and continuously tracked state. The state is continuously updated based on trial-level feedback because no external signal directly indicates its status, while the color cue changed trial-by-trial. These differences led us to predict dissociable neural substrates for encoding each input.

2.2.1.1 State Belief Decoding: Prior research suggests the orbitofrontal cortex (OFC) encodes state information (8-10). We hypothesized that OFC activity would correlate with model-derived trial-level state beliefs (Fig. **5A**). Following cue onset, we found significant clusters in the parahippocampus, fusiform gyrus, middle frontal gyrus (MFG), and visual association cortex

(Fig. **5B**). Probe onset data revealed clusters in the visual association cortex, middle temporal gyrus (MTG), primary sensory cortex, and superior parietal lobe (SPL). Although our predefined threshold (minimum cluster size of 21 voxels; $p < .01$; voxel $p < .005$) did not yield an OFC cluster, an exploratory analysis with a lower voxel-level threshold ($t = 2.412$) did identify a sub-threshold cluster in the OFC.

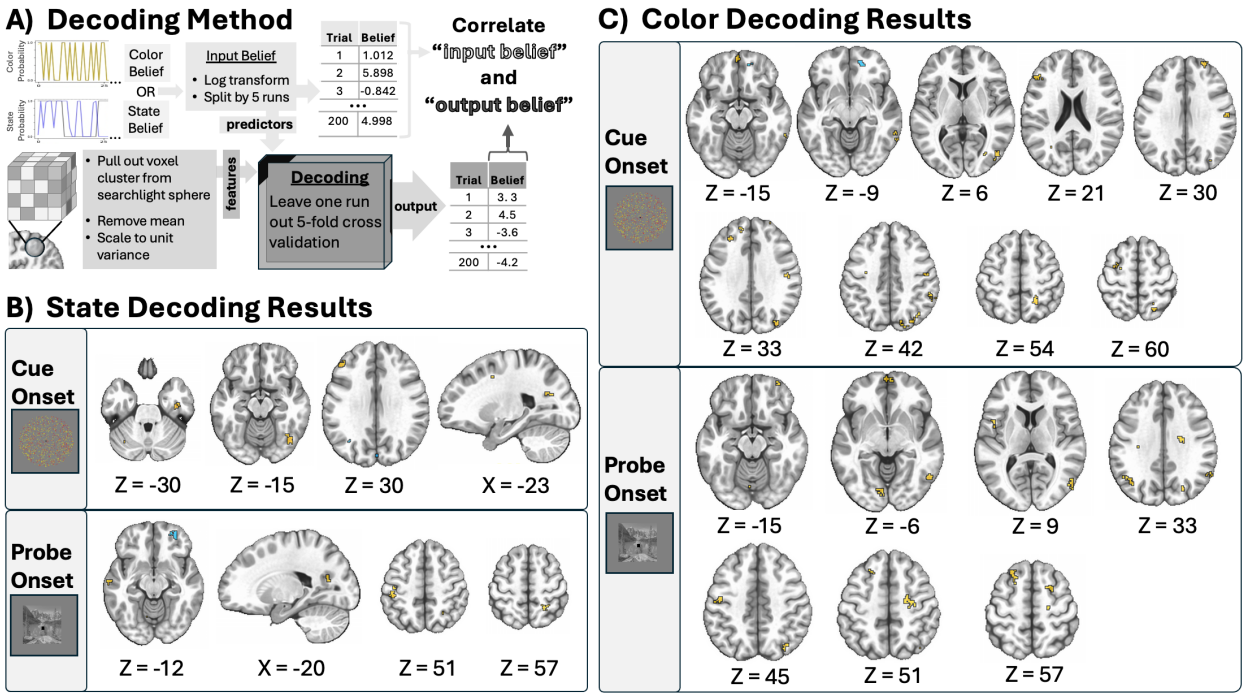

Fig 5. Input Decoding. (**A**) Depiction of decoding method for color and state belief. Decoding results for color belief (**C**) and state (**B**) belief with cue and probe epochs. Results based on minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005). The face stimuli in probe example is the corresponding author.

2.2.1.2 Color Belief Decoding: Previous studies showed that externally presented cues, linked to task rules, elicited activity in the frontoparietal and dorsal attention networks (11, 12). Given that the color belief was derived from externally presented perceptual information (the cue at the start of the task trial) and is linked to task rules (cue color informs the subsequent task-to-perform), we predicted that trial-level color beliefs would be decodable from regions typically found within the visual attention and frontoparietal networks, such as middle temporal visual area (MT), intraparietal sulcus (IPS), and lateral PFC. Consistent with this prediction, decoding analyses following cue onset revealed distributed clusters across prefrontal, temporal, limbic, occipital, and parietal cortices (Fig. 5C), including significant clusters in ventral attention (MT,

IPS) and frontoparietal regions (bilateral superior frontal gyrus, MFG, SPL). A significant cluster was also found in OFC. During probe onset, clusters emerged in inferior frontal cortex (including OFC and IFG), superior frontal cortex (MFG, preSMA, and primary motor cortex), and inferior temporal cortex (fusiform gyrus).

These findings suggest that external cues tied to task rules are processed in visual attention and frontoparietal networks. Moreover, these regions were distinct from regions encoding information about the non-observable state. This dissociation in neural representations of input information raises a critical question: how does the brain integrate these disparate representations to form a unified task output?

2.2.1.3 State and Color Uncertainty Parametric Modulation: Another critical question is how neural systems encode uncertainty of input representations. Specifically, it remains unclear whether regions tracking state and color beliefs modulate their activity in relation to the level of uncertainty inherent in these inputs. In other words, does higher uncertainty in these beliefs require greater neural activity to encode and maintain this information. To address this question, we examined the neural substrates representing uncertainty in state and color beliefs (Fig. **6A**).

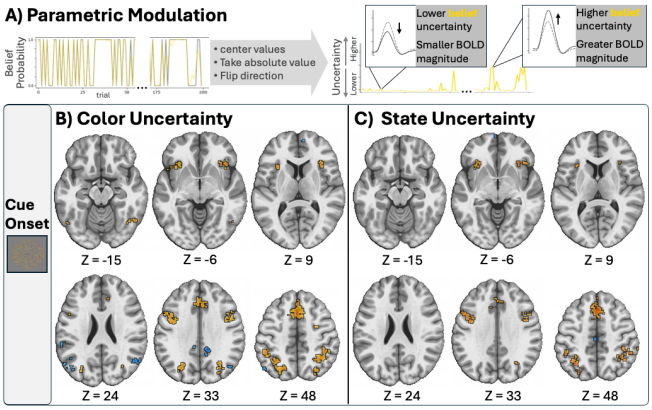

Fig 6. Input Parametric Modulation. (**A**) Depiction of parametric modulation analyses for color uncertainty and state uncertainty. Uncertainty calculated by centering color/state belief, taking the absolute value, and flipping the direction so that a larger value reflects greater uncertainty. Color (**B**) and state (**C**) uncertainty results based on cue epochs. Parametric modulation results for state and color uncertainty based on minimum cluster size of 23 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005).

We observed substantial overlap in brain regions that responded to both color (Fig. **6B**) and state uncertainty (Fig. **6C**). Specifically, MFG, medial prefrontal cortex (mPFC), IPS, as well as ventrolateral prefrontal cortex (vlPFC) all showed increased activity with greater uncertainty in both color and state beliefs. Moreover, there was a cluster in the inferior frontal sulcus (IFS) that overlapped between state decoding and state uncertainty parametric modulation as well as a cluster in the fusiform gyrus that overlapped between color decoding and color uncertainty parametric modulation. No other regions demonstrated such overlap between decoding and uncertainty parametric modulation.

2.2.2 Joint representation and the output from integration

In our framework of cognitive integration, the joint representation is defined as the posterior distribution formed by integrating multiple probabilistic inputs. This joint probability, which reflects specific combination of color and state, can be used to predict the output of integration, expressed as a task belief that determines which task to perform. Following cue onset, we expected preparatory activation in brain regions encoding task belief. We hypothesized that the joint probability and task belief would be decoded from frontoparietal regions associated with cognitive control, such as more caudal regions of MFG and IPS.

2.2.2.1 Joint Probability and Task Belief Decoding: In line with the above mentioned predictions, decoding analyses of joint probability identified significant clusters in MFG and IPS, as well as hippocampal regions and FFA/PPA (Fig. 7A). Robust clusters were also observed in the supplementary motor cortex regions across the primary motor and sensory cortices towards the IPS, with similar patterns during the probe period. Furthermore, for decoding task belief, distributed clusters were found in regions where the joint probability was decoded (Fig. 7B). Of particular note, task belief was decoded from the same regions spanning from supplementary motor cortex regions across the primary motor and sensory cortices into the IPS. This pattern may

reflect the fact that participants could preparatorily activate the left or right hand depending on which task they believed they should perform.

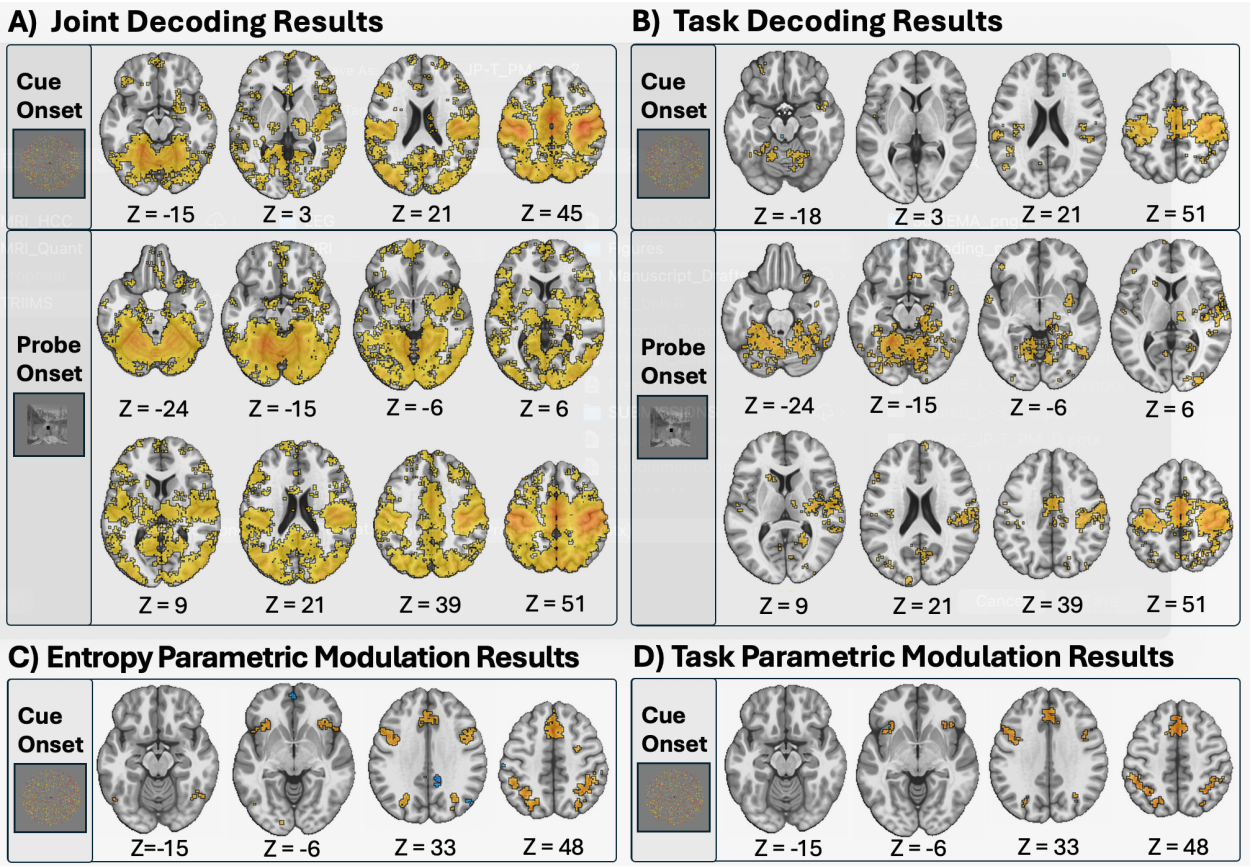

Fig 7. Integration and Output Decoding and Parametric Modulation. Decoding results (**A-B**) for joint probability (**A**) and task belief (**B**) with cue and probe epochs based on minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005). Parametric modulation results (**C-D**) for entropy (**C**) and task uncertainty (**D**) with cue epochs based on minimum cluster size of 23 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005). The face stimuli in the probe example is the corresponding author.

2.2.2.2 Joint Representation and Task Uncertainty: We further conducted a parametric modulation analysis to identify regions where activity scales with integrated uncertainty (entropy) and task uncertainty. Entropy captures the combined uncertainty of state and color inputs; entropy remains low when both are certain, high when both are uncertain, and intermediate otherwise. Consequently, areas modulating their activity with state and color uncertainty should also be sensitive to entropy. These regions should also modulate their activity with task uncertainty given that the joint probability and task belief were decoded from overlapping regions. Consistent with

these predictions, following cue onset, regions including MFG, mPFC, IPS, SFG, and vlPFC exhibited significant modulation with both entropy and task uncertainty (Fig. **7C-D**).

In sum, frontoparietal regions, such as MFG and IPS, previously found to display hub properties (12), have thus far shown two key findings. First, their activity fluctuated with measures of input uncertainty, integrated uncertainty (entropy), and output uncertainty. Second, these regions showed strong decoding for both the joint probability (representing the integrated inputs) and the task belief (output). One interpretation based on these observations is that these regions integrate uncertainty from task-relevant information sources and utilize this signal to bias the integration output. To test this interpretation, we performed a direct analysis of functional connectivity between regions encoding the inputs, the joint uncertainty, and the task outputs.

2.2.3 Integrated uncertainty interactions with functional connectivity

Briefly, we conducted beta-series connectivity (13) analyses between 400 cortical ROIs (14) using single-trial beta estimates from all trials across cue and probe epochs. For each pair of ROIs, we ran a mixed-effects regression model in which the beta value from one ROI (defined as "target ROI") was predicted by that from another (defined as "source ROI"), with entropy as a moderator. This yielded estimates for both the main effect of connectivity and its interaction with the integrated uncertainty. We focused on the interaction effect for the purpose of this analysis. The interaction effect indicates that the connectivity between brain regions is not static but varies according to the level of integrated uncertainty. In other words, the functional connectivity between the source ROI and target ROI fluctuates with the entropy of the joint representation. During the cue epoch, we predicted stronger coupling between regions encoding entropy and integration inputs (color and state) with higher uncertainty; suggesting that these networks dynamically adjust their communication to better integrate information. Similarly, during the probe epoch, we predicted that higher uncertainty would modulate the coupling between regions

39 encoding the task output and entropy; reflecting networks dynamically adjusting communication
40 to select and execute this selected task-to-perform.

41 To test these predictions, we employed a regression modeling framework using the
42 observed interaction coefficient as our dependent variable, while incorporating region-specific
43 encoding strength estimates (t-values from the whole-brain decoding analyses) for entropy, task,
44 color, and state as predictors (Fig. **8G**). This approach quantified how the connectivity interaction
45 with integrated uncertainty (i.e., entropy) is modulated by the encoding strength of various task-
46 related representations (i.e., entropy, task, color, state) across regions. Connectivity analyses were
47 implemented in a fully data-driven manner across all 400 ROIs to ensure an unbiased exploration
48 of the distributed network dynamics underlying cognitive integration.

49 Both cue and probe epochs exhibited distinct brain-wide connectivity patterns (Fig. **8A-**
50 **B**). During the cue epoch, higher integrated uncertainty was associated with increased
51 connectivity between source entropy and target color (t=3.787, p<.001), and decreased
52 connectivity between source entropy and target task (t=-5.665, p<.001 (Fig. **8D-E**). This pattern
53 suggests that increased uncertainty enhances external sensory processing while reducing
54 interaction with task output. In contrast, during the probe epoch, higher integrated uncertainty
55 significantly enhanced connectivity with target task regions (t=4.176, p<.001; Fig. **8B**), indicating
56 that integrated uncertainty primarily influences task-related connectivity as decisions are being
57 executed. Moreover, using source entropy from the cue epoch to predict target task from the
58 probe epoch revealed a significant positive modulation (t=6.220, p<.001 (Fig. **8C-F**), indicating
59 top-down influences from earlier cue periods shape task-related output at probe onset. Given that
60 entropy was most robustly decoded from regions such as MFG and IPS, which tend to exhibit
61 hub-like properties (12), these results further suggest a role for frontoparietal hub regions more
62 specifically in encoding integrated uncertainty during integration during cue periods and biasing
63 downstream task regions at probe onset. These findings collectively suggest that as integrated

uncertainty increases, connectivity patterns across brain regions dynamically adjust to bias the system toward the correct task output. Further supporting an account of integration where frontoparietal hub regions bias activity patterns to influence the task output for adaptive behavior.

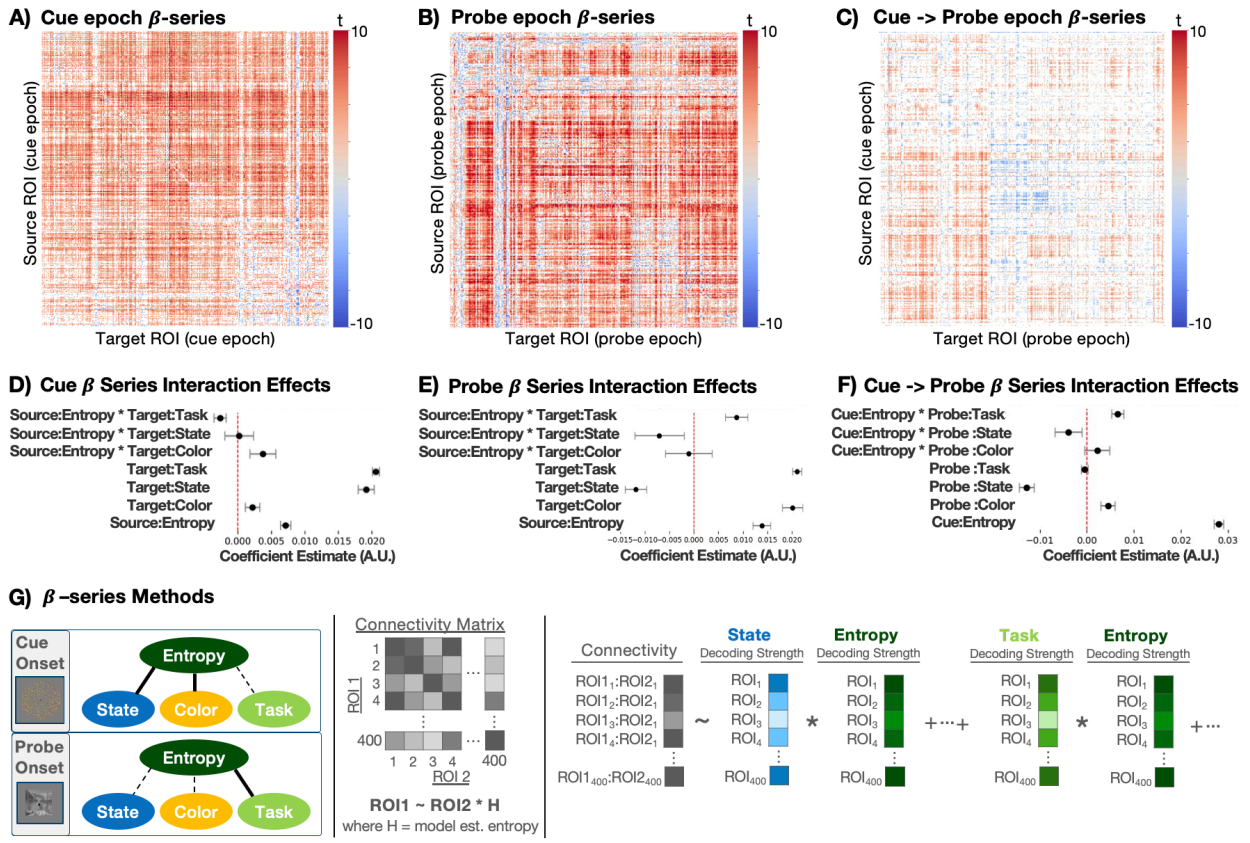

Fig 8. Functional Connectivity and Integration. Beta-series connectivity matrices of entropy interaction for the cue (**A**) and probe (**B**) epochs as well as cue-to-probe (**C**) epochs, displaying T-values from a mixed-effects regression analysis. Each cell represents the effect of a "source" ROI (rows) on a "target" ROI (columns). Warmer colors indicate higher positive T-values, while cooler colors indicate more negative T-values. **(D-E)** Forest plot illustrating the significant interaction effects (coefficient estimates ± 95% CI) between source ROI entropy and target ROI encoding of task, state, or color. Associated statistic tables can be found in the supplementary tables 1-3. During the cue epoch (**A**), higher source entropy positively modulates connectivity with target color encoding but negatively modulates connectivity with target task encoding, suggesting a bias toward external sensory processing under uncertainty. In contrast, during the probe epoch (**B**), higher source entropy significantly modulates connectivity with target task encoding, indicating a shift toward task-related output processing as decisions are executed. Finally, when using cue epochs as the source and probe epochs as the target (**C**), higher source entropy (at cue epochs) significantly modulates connectivity with target task encoding (at probe epochs), suggesting a directional influence from cue to probe where integrated entropy at cue, when integration occurs, influences task-related output processing at probe, when decisions are executed. (**G**) Predictions for connectivity patterns following cue versus probe where a thick solid line indicates stronger connectivity and a thin dashed line indicates weaker connectivity; Depiction of connectivity matrix and model regressors used for to fit the observed connectivity matrix. The face stimulus in the probe example is of the corresponding author.

2.2.4 Functional hubs and uncertainty

Our integration framework predicts that regions that integrate internal and external information—such as the MFG and IPS—should exhibit hub-like properties. Specifically, these regions should dynamically track the integrated uncertainty of task-relevant inputs and convert this uncertainty into a biasing signal that guides task output. To verify this prediction, we computed voxel-wise participation coefficient (PC) scores using a previously developed method (15). PC estimates ranged from 0-1 and hub regions were a priori defined as having a PC > 0.65. Based on these criteria, MFG and IPS are hub regions in the present study (Fig. 9). Moreover, there is a strong overlap between regions fluctuating with uncertainty and hub regions. These findings further support our account in which hub regions integrate uncertainty from diverse information sources to bias task output.

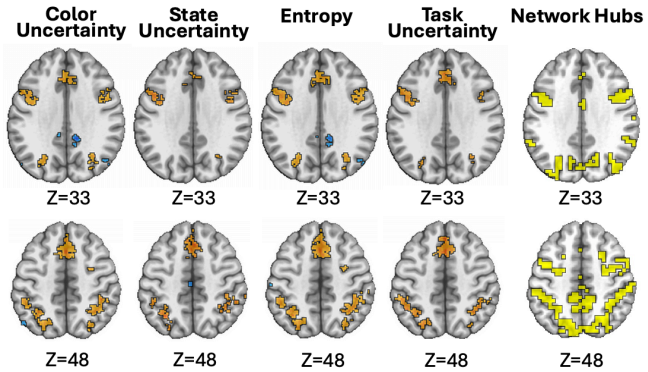

Fig 9. Network Hubs. Network Hub regions compared to regions fluctuating with uncertainty/entropy. Strong overlap between hub regions, such as MFG and IPS, and regions fluctuating with uncertainty/entropy. Network hubs defined as regions with a PC value > 0.65.

2.2.5 Updating belief representations

Previous sections have focused on how the integration of information biases task output. However, our account of integration would be incomplete without understanding how input representations are updated following feedback. In our model, the state belief—unlike color belief, which is updated trial-by-trial after cue onset—requires continuous monitoring across trials

and updating because no direct external signal indicates state. Updating state belief thus involves evaluating: (1) the relationship between the received feedback and the performed task, (2) the joint representation of state and color beliefs that led to the task decision, and (3) the current state belief. Thus, state updating required multiple information sources to generate prediction errors and update state belief.

To assess updating processes, we first investigated brain regions where activity was modulated by both task and state prediction errors following feedback (Fig. **10A**). Presumably, incorrect feedback first triggers a task prediction error, which will then trigger a re-evaluation of the integrated product and the inputs to integration (color and state belief). For instance, if a subject is highly confident in their color belief yet receives negative feedback, a state prediction error is expected as the system adjusts the state belief. Additionally, we examined regions where activity varied with the trial-to-trial change in state belief, termed the state derivative (Fig. **10D**). These state derivative analyses were time-locked to the cue onset on trial N to investigate updating processes occurring following feedback on the previous trial (trial N-1). Due to design limitations, we were unable to dissociate color updating from initial encoding.

2.2.5.1 Prediction Errors: We focused on state and task prediction errors while controlling for general error processing (i.e., whether feedback was correct). Using feedback-locked data, we predicted that mPFC voxels would show increased BOLD responses with larger prediction errors, consistent with its known role in conflict and error monitoring (10, 16-19). As expected, both state and task prediction errors showed greater activation in mPFC (Fig. **10B-C**). Beyond these expected regions, we also found that regions in the right inferior frontal junction (IFJ), posterior parietal cortex (PPC), lateral PFC, and subcortical structures, such as the thalamus and basal ganglia, showed a greater magnitude in their BOLD response with greater state and task prediction errors.

2.2.5.2 State Derivative: We expected greater changes in state belief from trial N-1 to trial N (i.e., a higher state derivative) to be associated with increased activity in the OFC, a region implicated in tracking state information (8-10). Furthermore, given that subcortical structures such as the basal ganglia are involved in working memory updating, we anticipated these regions would also exhibit stronger BOLD activity with larger state derivatives, reflecting greater updating demands. In line with these predictions, both the OFC and basal ganglia (caudate nucleus), as well as the medial thalamus, showed increased BOLD magnitude with greater state derivative (Fig. **10E**).

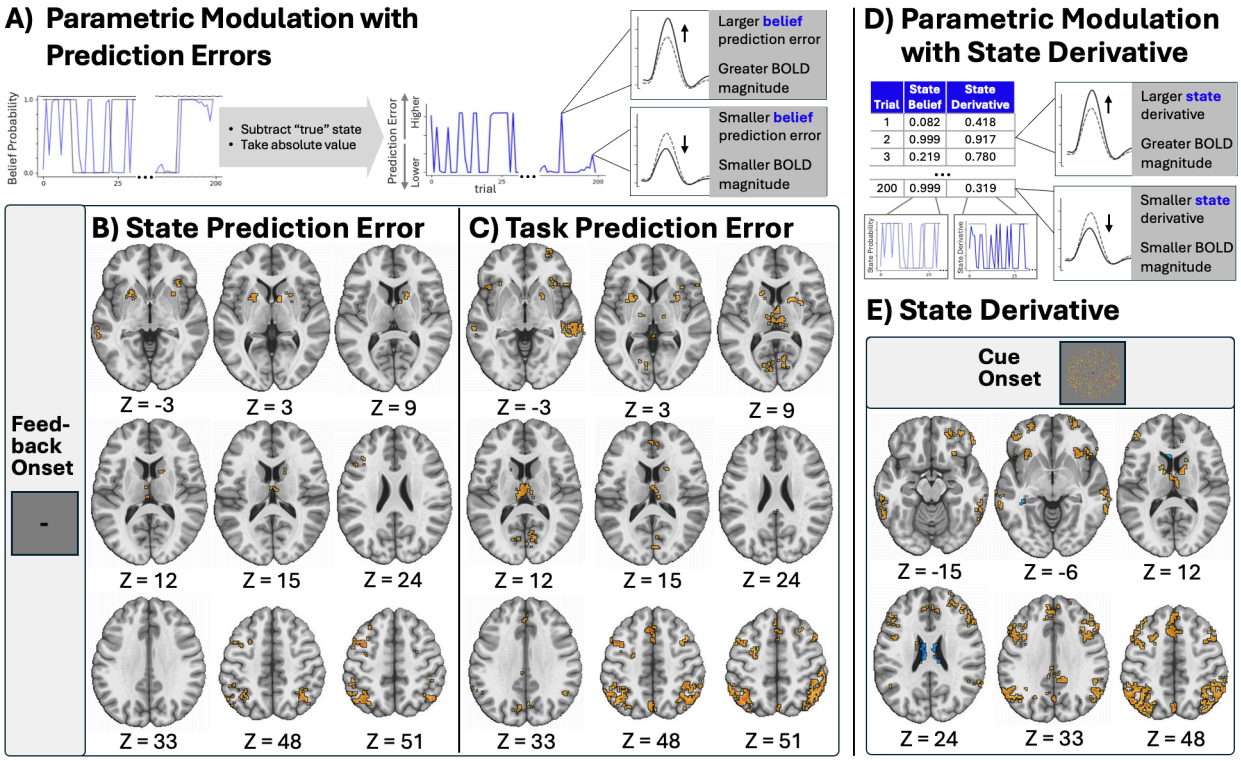

Fig 10. Prediction Errors and Updating Parametric Modulation. (**A**) Depiction of parametric modulation analyses with state and task prediction errors, and state derivative. Prediction error calculated by subtracting the "true" state/task from the model estimated state/take and taking the absolute values. Parametric modulation results with state (**B**) and task (**C**) prediction errors based on feedback epochs. (**D**) State derivative calculated by subtracting the state belief from trial N-1 from the state belief on trial N. (**E**) Parametric modulation results with state derivative based on cue epochs. Results based on minimum cluster size of 23 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005).

2.2.6 Functional hubs and updating

Given that frontoparietal hub regions (i.e., IPS) showed a greater BOLD magnitude with greater prediction errors and state derivative, we conducted additional functional connectivity analyses to examine how entropy influences prediction error processing. Because prediction errors occur only when both the expected outcome and its associated uncertainty are known, we hypothesized that higher entropy at probe would modulate prediction error processes at feedback.

We calculated connectivity using the same method in section 2.3.3 with a few key differences. Connectivity was calculated two ways: (1) between probe and feedback epochs, to assess how entropy during task execution (probe) affects prediction errors at feedback, and (2) solely within the feedback epoch. We predicted that entropy-modulated connectivity between the probe and feedback epochs would be positively correlated with prediction errors, indicating that these networks dynamically adjust their interactions to signal the need for information updates.

We found that higher integrated uncertainty, i.e. entropy, was associated with a significant positive modulation of functional connectivity between source entropy (from probe epochs) and state prediction error (from feedback), t=2.057, p=.040, and a significant negative modulation of functional connectivity between source entropy (from probe epochs) and task prediction error (from feedback), t=-10.922, p<.001 (Figure **11A-B**). In contrast, when connectivity was computed using only feedback epochs, no significant modulation was observed for state prediction errors, t=-0.856, p=.392, or task prediction errors, t=0.361, p=.718 (Figure **11C-D**). Thereby suggesting that integrated uncertainty at task execution specifically modulates state prediction error processes following feedback. In other words, when integrated uncertainty is high, meaning both color and state are highly uncertain, there is enhanced coupling between regions encoding integrated uncertainty and regions fluctuating their activity profiles with state prediction errors. In parallel, there is reduced coupling with regions fluctuating their activity profiles with more general task prediction error processes. Similar to how frontoparietal hub regions encoding integrated

uncertainty drive task output processes, these results suggest that these same frontoparietal hub regions leverage integrated uncertainty to bias state prediction error processing under high uncertainty conditions. This drives targeted updating of the state belief, which needs to be tracked and updated on a trial-by-trial basis.

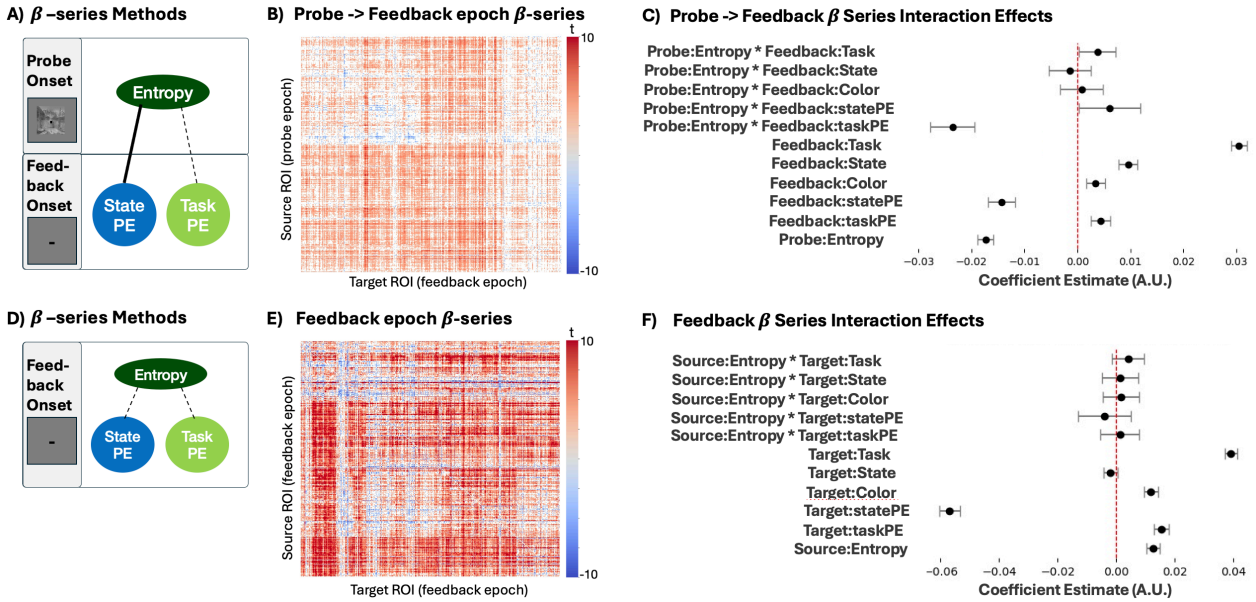

Fig 11. Functional Connectivity and Prediction Errors. Predictions for connectivity patterns (**A&D**) where a thick solid line indicates stronger connectivity and a thin dashed line indicates weaker connectivity. Beta-series connectivity matrices of entropy interaction for the probe-to-feedback (**B**) epochs and feedback (**E**) epochs, displaying T-values from a mixed-effects regression analysis. Each cell represents the effect of a "source" ROI (rows) on a "target" ROI (columns). Warmer colors indicate higher positive T-values, while cooler colors indicate more negative T-values. **(C&F)** Forest plot illustrating the significant interaction effects (coefficient estimates ± 95% CI) between source ROI entropy and target ROI encoding of task PE or state PE. Associated statistic tables can be found in the supplement tables 4-5. (**B**) When using probe epochs as the source and feedback epochs as the target, higher source entropy (at probe epochs) significantly modulates connectivity with target state prediction error modulation (at feedback epochs), suggesting a directional influence from probe to feedback where integrated entropy at probe, when decisions are executed, influences task-related state belief error processing at feedback, when state belief can be evaluated and updated. (**E**) When using only feedback epochs there were no significant interactions between entropy and prediction errors.

Discussion

Our primary goal was to determine the neurocognitive mechanisms supporting cognitive integration. Specifically, how uncertain non-observable beliefs/expectations are integrated with observable perceptual information to guide behavior. To address this, we developed a Bayesian

model capturing the full transformation from inputs to outputs by specifying: (1) the representational structure of inputs, (2) the integrated representation (joint probability distribution) encoding the integrated uncertainty (entropy) of these inputs, (3) the probabilistic output (task) belief, and (4) the feedback-recruited prediction errors and updating mechanisms.

The first two aims identified the representational structure of the inputs, integrated representation, and output task belief. Our first aim identified the neural substrates representing the to-be-integrated information. Probabilistic decoding revealed perceptual inputs and internally maintained information were encoded in largely distinct neural substrates. Color belief was primarily represented in regions typically associated with the FPN, DAN, and VAN (11, 12), while state belief was primarily represented in regions typically associated with the DMN (9, 20, 21). The fact that distinct regions encoded color and state belief further underscored the importance of investigating the integration of these input sources of information. Our second aim identified the neural substrates representing the integrated representation and output task belief. Multinomial logistic regression decoding analyses revealed that the specific combination of inputs— captured by a joint probability distribution—accurately predicted the state-color integration and task output. Moreover, following probe onset, there were clusters in sensory-motor regions where the joint probability was decoded from that overlapped with regions where state or color (inputs) and task (output) beliefs were decoded.

Our third aim examined how integration transforms distinct inputs into a single, unified task (output) representation. Parametric modulation analyses revealed similar sets of regions modulate their activity with individual input/output uncertainty and integrated uncertainty (entropy). Notably, this set of regions included frontoparietal regions such as middle frontal gyrus (MFG) and intraparietal sulcus (IPS)—hub regions connected to multiple large-scale networks (3, 12, 22, 23). Given that these regions exhibit strong network hub properties, we further utilized connectivity analyses to test the role of FPN hubs in integration. During integration, increased

entropy enhanced connectivity with regions encoding external sensory information while reducing connectivity with task-output regions. Conversely, during task execution, higher entropy primarily enhanced connectivity with task-output regions. Temporally directional cue-to-probe connectivity analyses confirmed entropy influenced task-output connectivity. Thus, entropy appeared to influence task selection. These results suggest FPN functions as a low dimensional readout system that detects and combines input uncertainty and then resolves output competition arising from this integrated uncertainty. The integrated uncertainty and integrated representations encoded in FPN hubs may be converted into a biasing signal that modulates downstream task-relevant regions and drives the output of a single, unified task belief (24, 25).

Our final aim investigated how integrated uncertainty (entropy) further impacted prediction errors and updating processes. Results indicated that updating processes recruited subcortical structures, such as the basal ganglia and the mediodorsal thalamus, as well as frontal and parietal regions, including the IFG, MFG, SFG, and IPS/PPC. Some of these regions (IFG, MFG, IPS) overlapped with regions that fluctuated their activity patterns with prediction errors, entropy, and color/state uncertainty. Moreover, temporally directional probe-feedback connectivity analyses found that entropy influenced prediction errors; suggesting the updating system relies on a biasing signal similar to the integration system. In other words, FPN hubs utilize integrated uncertainty to bias regions involved in the detection of violations in expectations (prediction errors) to drive targeted updating of the internally maintained state belief.

Our findings highlight the importance of uncertainty in integration and task control. Sensory research has long demonstrated that neural activity encodes information probabilistically rather than in an all-or-none fashion (26, 27). However, decision making requires a winner-take-all process where only the most probable option is selected (28). Our framework proposes that probabilistically encoded inputs are integrated into a joint distribution and generate a conditional probability from which the most likely action is selected. Our results suggest that the human brain

leverages the probabilistic information encoded in these sensory inputs and processes them within a common set of frontoparietal regions as opposed to resolving uncertainty at the input level and considering only the categorical information during integration. Moreover, although the output belief drives motor plans in a winner-take-all manner, this output belief also encodes uncertainty to facilitate subsequent updating.

Furthermore, these results suggest regions modulating their activity with input/output uncertainty and integrated uncertainty reflect a core function of cognitive integration. In our framework, the entropy of the joint distribution serves as a critical signal that detects competition, noise, and variability across multiple task-relevant input sources. This signal enables the selection of the most likely correct action plan and decision (i.e., task) based on input information. Recent work by Li, Sprague (29) similarly indicates that uncertainty signals guide targeted prioritization of multiple representations in working memory, resolving competition between representations. We found similar results with prediction errors and updating processes. Greater entropy enhanced connectivity related to state prediction errors, driving targeted updating of internally maintained state information. Therefore, the maintenance of probabilistic representations across diverse brain regions and their integrated uncertainty is integral to both the continuous integration of this information, facilitating adaptive behavior, and targeted updating of this information.

In summary, this work demonstrates how the frontoparietal system actively integrates internal and external information to guide adaptive behavior. By defining the representational structure of the inputs, the integrated product (joint distribution), and the output (task belief)—along with their associated uncertainty measures—we were able to track the representational structure and transformation of information from initial encoding through integration out to action selection. Results revealed that internal and external inputs were represented in distinct brain structures before converging in FPN hub regions that dynamically modulated connectivity based on integrated uncertainty estimates. The task output was primarily encoded in premotor and motor

areas, reflecting the final selection process. Finally, updating mechanisms recruited both FPN and subcortical structures (basal ganglia, thalamus) to update internal representations based on feedback. This work answers a fundamental question of how the human FPN and neurocognitive systems integrate and update task-relevant information for adaptive, goal-directed behavior.

Materials and methods

3.1 Experimental design

To address cognitive integration, we utilized an experimental design in which participants integrated one external and one internal source of information to determine the correct task for each trial (Fig. **2**). We developed a Bayesian model (Fig. **3A**) to capture computational processes underlying cognitive integration in this experimental design. Participants completed an fMRI scanning session while performing this cognitive integration task. Before scanning, all participants completed a tutorial and practice session. Scanning began after participants achieved a minimum practice score of 80%. We collected five functional runs, each containing 40 task trials, resulting in a total of 200 trials across the session. Each functional run lasted approximately 8.5 minutes.

3.2 Participants

We recruited 49 healthy participants (19 male, M_{age} = 22.396, SD_{age} = 4.372, $Range_{age}$ = 18-35 years) from the University of Iowa and surrounding area to participate in this study. All participants had normal or corrected to normal visual acuity and color vision, were right-handed, and reported no history of epilepsy, psychiatric, or neurological conditions. This study was approved by the University of Iowa Institutional Review Board and conducted in accordance with the principles outlined in the Declaration of Helsinki. All participants gave written informed consent. Out of these 49 participants, 11 were excluded for excessive movement and/or chance-

level behavioral performance during fMRI scanning, resulting in a final sample of 38 participants (15 male, M_{age} = 22.95, SD_{age} = 5.88, $Range_{age}$ = 18-35 years).

3.3 Experimental tasks

Each trial (Fig. **2B**) began with the presentation of an array of red and yellow dots moving randomly in a circle (cue screen; presented for 1000-milliseconds). During this presentation, participants determined if there were more red or yellow dots in this array. After a jittered delay (average duration of 3.75 seconds; range of 0.5 – 10.5 seconds), a probe screen appeared for 1500 milliseconds, showing a partially transparent face overlaid on a partially transparent scene image. At this point participants performed either the face task (report whether the face was male or female; Fig. **2D**) or the scene task (report whether the scene depicted a city or a natural landscape; Fig. **2D**). As previously described, participants needed to integrate the hidden state and perceptual color information to determine which task to perform. In one hidden state, more red dots indicated that the correct task was the face task, while more yellow dots indicated the scene task. In the other hidden state, this pairing was reversed: more red dots indicated the scene task, and more yellow dots indicated the face task. After a second jittered delay (average duration of 3.75 seconds; range of 0.5 – 10.5 seconds), a feedback screen appeared for 500 milliseconds, showing either a plus or minus sign. A plus sign indicated that the participant performed the correct task and made the correct discrimination judgement (male/female or city/landscape), while a minus sign indicated either the participant performed the incorrect task and/or made the wrong discrimination judgement. Finally, each trial ended with a jittered inter-trial-interval (average duration of 3.75 seconds; range of 0.5 – 10.5 seconds).

3.4 Computational model

In the Bayesian model, for trial i, the probabilistic state belief (S_i, representing the two color-task mappings in Fig. 2A) and probabilistic color belief (C_i) are integrated into a 2x2 joint distribution, which was used to infer the probabilistic task belief (T_i).

$$p(T_i \mid C_i, S_i) \quad (1)$$

Here, state belief (S_i), color belief (C_i), and task belief (T_i) are latent variables and state belief (S_i) and color belief (C_i) are inferred from the directly observable color proportion (D_i) and response made by the participant (R_i). Because the previous state belief (S_{i-1}) should influence the current state belief (S_i), which should in turn influence the subsequent state belief (S_{i+1}), a trial-wise diffusion of S_i was implemented with the following equation.

$$P(S_i) = P(S_i) * (1 - \theta_D) + 0.5 * \theta_D \quad (2)$$

In other words, at the beginning of each trial, a confusion in the form of random guess (i.e., 0.5 probability) of the state is introduced to the previous state belief with a weight of θ_D, which ranges from 0 (i.e., no confusion) to 0.5 (30, 31).

Color belief (C_i) was estimated from the observable color proportion (D_i) using a sigmoid function. The free parameters (θs) estimating this sigmoid function determined the intercept (accounting for color bias) and slope (accounting for color uncertainty/perception).

$$C_i = p(R_i \mid \theta s,\ D_i) \quad (3)$$

The equation $p(T_i \mid C_i, S_i)$ was combined with two additional free parameters, the participants respective error rates for the face and the scene tasks, to predict the responses (R_i). R_i was coded as 0 for correct task and correct response or 1, which could mean either correct task but incorrect response or incorrect task and either correct or incorrect response. By using the trial-wise values of the directly observable variables (D_i and R_i), the generative model was inverted to infer latent variable distributions (Fig. 3A).

The proposed model assumes probabilistic representations for both state belief (S_i) and color belief (C_i). As a result, the 2x2 joint probability distribution encoded the combined uncertainty of these two sources of inputs. Furthermore, properties of this joint distribution, such

as entropy or uncertainty, were used in neuroimaging analyses to determine neural substrates of cognitive integration. Shannon's entropy was calculated using the following (Equation 4):

$$H = -1 * \Sigma\, p_i * \log(p_i) \tag{4}$$

Here, p_i refers to each cell in the 2x2 joint distribution. Shannon's entropy was used to quantify the expected uncertainty across the four possible outcomes in this joint distribution.

Uncertainty estimates were extracted from these latent variables and the joint distribution in order to more fully capture the representational structure. This included the uncertainty of the information that acted as the inputs for integration (state and color beliefs), the integrated representation (entropy of the joint distribution), the uncertainty of the output from the integrated representation (task belief), and the updating mechanisms (prediction errors and derivatives).

We extracted uncertainty from each of the three latent variables (state belief, S_i; color belief, C_i; task belief, T_i) using the following steps. First, each belief was centered by subtracting 0.5 from the values. Next, the absolute value of the centered beliefs was calculated to represent the uncertainty in the belief. Finally, for ease of interpretation, the signs of these uncertainty scores were inverted so that values closer to zero indicated lower uncertainty, whereas higher values indicated greater uncertainty. This calculation better aligned uncertainty scores with entropy estimates, where values closer to zero represented lower uncertainty in the expected outcome, and higher values represented greater uncertainty in the expected outcome. Prediction error was calculated by taking the difference between the model estimated belief and the true value. Finally, state derivative was calculated as the magnitude of the change from the state belief on trial N-1 to trial N. For example, if the state belief on trial N-1 was 0.83 and trial N was 0.99, then the state derivative would be 0.16 for trial N.

3.5 Computational model comparisons

This computational model assumes probabilistic representations for both state belief (S_i) and color belief (C_i), which are integrated into a joint distribution to determine the task belief (T_i). To test its validity, we compared this model against three alternative models: (1) a model with dichotomous state belief (S_i) and probabilistic color belief (C_i), (2) a model with probabilistic state belief (S_i) and dichotomous color belief (C_i), and (3) a model in which both state belief (S_i) and color belief (C_i) are dichotomous. Here, dichotomous belief refers to the fact that the information is represented in an all-or-none fashion that does not consider uncertainty. The models were evaluated using two criteria. First, we compared their ability to correctly predict the task performed by participants by looking at their prediction errors.

3.6 fMRI data acquisition

Imaging data was collected at the Magnetic Resonance Research Facility at the University of Iowa using a 3T GE SIGNA Premier scanner with a 48-channel head coil. Structural images were acquired using a multi-echo MPRAGE sequence (TR = 2348.52ms; TE = 2.968ms; flip angle = 8°; field of view = 256*256; 200 sagittal slices; voxel size = $1mm^3$). Functional images were acquired using an echo-planar sequence sensitive to blood oxygenated level-dependent (BOLD) contrast (multiband factor = 2; TR = 2039ms; TE = 30ms; flip angle = 75°; voxel size = $2mm^3$).

3.7 fMRI data preprocessing

All fMRI data were preprocessed using fMRIPrep version 23.2.0 (32) to reduce noise and transform data from subject native space to the ICBM 152 Nonlinear Asymmetrical template version 2009c for group analysis (33). Preprocessing steps include bias field correction, skull-stripping, co-registration between functional and structural images, tissue segmentation, motion correction, and spatial normalization to the standard space. As part of preprocessing, we obtained nuisance regressors including rigid-body motion estimates, cerebral spinal fluid (CSF), and white

99 matter (WM) noise components from the component-based noise correction procedure (34).
00 These nuisance regressors were entered in the regression model to reduce the influences from
01 noise and artifacts. We did not perform any spatial smoothing when preprocessing the data.

02 3.8 Statistical analyses

03 3.8.1 Mixed effects models predicting response times

04 To examine how trial-by-trial variations in cognitive integration influence behavior, we
05 conducted two separate mixed effects models predicting participant response times (RTs). The
06 models used either (1) the entropy of the joint distribution or (2) the uncertainty of the task belief,
07 which was the output of the joint distribution. Both models included random intercepts for
08 subjects. Response times faster than 200 milliseconds and slower than 3500 milliseconds (offset
09 of probe image) were excluded from analyses. The first trial of each block was also excluded. The
10 entropy of the joint distribution was winsorized to address outliers (values in the bottom 5% and
11 top 95% were replaced with the values at 5th or 95th percentile, respectively). Winsorized entropy
12 and task uncertainty values were approximately normally distributed and homoscedastic. We
13 performed two separate mixed effects regressions due to the high correlation between entropy and
14 task uncertainty ($r = .76$). We predicted that both greater entropy and greater task uncertainty
15 would predict slower RTs.

16 3.8.2 Parametric modulation

17 To relate our model parameters to fMRI data, we conducted parametric modulation to
18 identify brain regions where BOLD response amplitudes vary systematically with key parameters
19 from our computational model: (1) the uncertainty of the model predicted state belief, color
20 belief, and task belief, (2) prediction errors and derivatives of task and state beliefs, and (3) the
21 entropy of the joint distribution. These analyses aimed to capture the representational structure, its
22 uncertainty estimates, and computational processes underlying cognitive integration. This
23 approach provided insights into the neural substrates supporting cognitive integration.

Parametric trial-level model estimates for state, color, and task belief uncertainties were calculated with the procedures described in section 2.4. Parametric trial-level model estimates for prediction error were calculated by subtracting the true task or state value from each model-estimated state or task belief and taking the absolute value. This provided a continuous measure of prediction error, reflecting how much the participant's belief deviated from the true state or task. Larger values indicated greater prediction error, i.e., a greater violation of expectations. Note, this continuous measure of prediction errors was applied to both correct and incorrect trials. Here, a small prediction error reflected a relatively small amount of uncertainty in the state or task belief paired with a correct response, which would reflect a smaller discrepancy between expectations and outcomes. A large prediction error reflected a relatively small amount of uncertainty in the state or task belief paired with an incorrect response, which would reflect a large discrepancy between expectations and outcomes. A moderate sized prediction error reflected a larger amount of uncertainty paired with either a correct or incorrect response, which would reflect a moderate discrepancy between expectations and outcomes. Parametric trial-level model estimates for the state derivative were calculated by subtracting the state belief on the previous trial from the state belief on the current trial and then taking the absolute value. This provided a continuous measure of the magnitude of the change in state belief from trial-to-trial. Larger values indicated a greater change in state belief. In other words, a greater degree of belief updating from the previous trial to the current trial.

The entropy, uncertainty, prediction error, and derivative estimates were then included as parametric modulators in a mass univariate GLM using AFNI's 3dDeconvolve with a voxel-wise restricted maximum likelihood (REML) estimate of an ARMA (1,1) temporal correlation model using AFNI's 3dREMLfit (35). This approach allowed us to assess how each modulator (uncertainty, prediction error, derivative and entropy) modulates the hemodynamic response function (HRF) magnitude across brain regions on a trial-by-trial basis, beyond trial-averaged

effects from the event epochs. Finally, we conducted mass univariate t-tests on the resulting beta values to test whether they differed from zero across participants. These tests identified structures where changes in entropy, uncertainty, prediction error, or derivative estimates corresponded to significant changes in the BOLD response. To correct for multiple comparisons, we used 3dClustSim with the residuals from the REML estimate to estimate the minimum cluster size for a corrected cluster-level threshold. Significant clusters for parametric modulation are based on these calculated minimum cluster sizes of 23 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005).

3.8.3 Probabilistic decoding

We conducted decoding analyses to identify the neural substrates encoding the following. First, the input sources of information, state belief (S_i) and color belief (C_i). Second, the output of this integrated product, task belief (T_i).

We first obtained trial-by-trial estimates of BOLD response amplitudes using a least-square-all approach implemented in AFNI's 3dLSS (35). Decoding was performed via linear regression (Python package scikit-learn; (36)), with trial-level model estimated participant beliefs (state belief, S_i, color belief, C_i, or task belief, T_i) serving as the predicted variable, and the trial-level beta estimates from 3dLSS as features. This decoding analysis was conducted separately for three distinct 3dLSS analyses, each time-locked to a different task phase (cue onset, probe onset, and feedback onset).

Participant beliefs ranged from zero to one, where boundary values (i.e., near 0 or 1) indicated higher confidence (lower uncertainty), and values near 0.5 reflected lower confidence (greater uncertainty). To better meet the assumption of normality, these beliefs were log-transformed before fitting decoding, as their underlying trial-level distributions were bimodal, with most values near the boundaries. We applied a whole-brain searchlight decoding approach, with each spherical searchlight having an 8mm radius. Within each searchlight sphere, voxel

features were standardized by removing the mean and scaling to unit variance. For decoding, trial-wise betas were split into five groups based on the five functional runs, and a leave one group out cross validation approach was used to predict the trial-wise estimated beliefs. To assess decoding performance, we correlated the predicted trial-level belief values (decoding output) with the actual model estimated belief values (decoding input), identifying brain regions where the predicted values closely matched the true values (i.e., where decoding performance was higher). Finally, we conducted mass univariate t-tests that tested if these correlation values across subjects were significantly different from zero. To correct for multiple comparisons, we used 3dClustSim with the residuals from a voxel-wise restricted maximum likelihood (REML) estimate of an ARMA (1,1) temporal correlation to estimate the minimum cluster size. Significant clusters for decoding results are based on these calculated minimum cluster sizes of 21 voxels (cluster p-value threshold of 0.01 and a voxel p-value threshold of .005).

3.8.4 Joint probability decoding

Similar to probabilistic decoding (section 2.9.2), we first obtained trial-by-trial estimates of BOLD response amplitudes using a least-square-all approach implemented in AFNI's 3dLSS (35). We then conducted a multinomial logistic regression (Python package scikit-learn; (36)), with the four possible combinations of state and color belief serving as the four categorical predicted variables, and the beta estimates from 3dLSS as features. This analysis was also conducted separately for three distinct 3dLSS analyses, each time-locked to a different task phase (cue onset, probe onset, and feedback onset).

We applied a whole-brain searchlight decoding approach, with each spherical searchlight having an 8mm radius. Within each searchlight sphere, voxel features were standardized by removing the mean and scaling to unit variance. For decoding, trial-wise betas were split into five groups based on the five functional runs, and a leave one group out cross validation approach was used to predict the trial-wise estimated beliefs. Because there were four possible classes (state1-

99 color1, state1-color2, state2-color1, state2-color2), chance level decoding performance was set at
00 0.25. Finally, we conducted mass univariate t-tests that tested if these decoding accuracy values
01 across subjects were significantly different from 0.25. To correct for multiple comparisons, we
02 used 3dClustSim with the residuals from a voxel-wise REML estimate of an ARMA (1,1)
03 temporal correlation to estimate the minimum cluster size. Significant clusters for decoding
04 results are based on these calculated minimum cluster sizes of 21 voxels (cluster p-value
05 threshold of 0.01 and a voxel p-value threshold of .005).

06 3.8.5 Connectivity analysis

07 We conducted a customized connectivity analysis to examine task-related connectivity
08 patterns between ROIs during the cue, probe, and feedback epochs, and to assess how trial-by-
09 trial entropy values modulate these patterns. Our approach builds on the beta-series connectivity
10 framework (13) by first extracting single-trial beta estimates from LSS deconvolution using the
11 Schaefer 400-ROI atlas (14), and adapted the generalized psychophysiology interaction
12 framework by fitting a regression model to the trial-wise beta values (37). Note that this same set
13 of LSS single-trial beta estimates were used for the decoding analyses.

14 For each subject, we computed pairwise regression models for all ROI pairs (excluding
15 self-coupling). In these models, the trial-wise beta values from a "target" ROI was predicted by
16 the beta values from a "source" ROI, and its interaction with the trial-by-trial entropy estimates.
17 To control for outliers, independent variables were winsorized and centered before interaction
18 terms were computed. We then fitted ordinary least squares regression models within subjects,
19 and combined these results in a mixed-effects framework across subjects, using random intercepts
20 to account for subject variability. This regression approach yielded estimates of both the main
21 connectivity effect and its interaction with integrated uncertainty. The interaction effect indicates
22 how the functional connectivity of one ROI on another ROI varies with the entropy of the joint
23 representation, and is thus the main focus of this analysis. Group-level statistical tests (e.g., one-

sample t-tests) were subsequently applied to the coefficient estimates of interaction effects, and connectivity matrices were visualized as heatmaps, thresholded by FDR $q < 0.05$.

The above analysis yielded a 400 by 400 connectivity matrix for the cue and probe epochs. In the second stage of the connectivity analysis, we then sought to determine how representations encoded by different ROIs may have contributed to the observed connectivity pattern. Specifically, we predicted the observed functional connectivity patterns using a model that included the representational strength of source and target ROIs (e.g., entropy, color, state, and task) and their interactions (e.g., source ROI's entropy interaction with target ROI's task belief). Conceptually, this interaction term captures how the relationship between a target ROI's task encoding and connectivity strength is moderated by the source ROI's entropy encoding. In other words, if a source region encodes high uncertainty, its influence on connectivity may depend on whether the target region strongly encodes the relevant task information. We fitted an ordinary least squares regression model using these predictors. All in all, this procedure yielded coefficient estimates that capture how the encoding strengths of different ROIs can predict the observed functional connectivity patterns.

3.8.6 Functional network hub analysis

To assess the connector hub-like properties of regions fluctuating with entropy/uncertainty, we calculated participation coefficient (PC) values, which measure how diversely connect a node is within a network (38, 39). This result is implemented using data and methods developed in Reber, Hwang (15), and briefly described here.

The PC value for each ROI, PC_i is defined as

$$\mathrm{PC_i} = 1 - \sum_{s=1}^{N_M} \left(\frac{k_{is}}{k_i}\right)^2 \quad (5)$$

where k_i is the sum of the total functional connectivity weight for ROI i, k_{is} is the sum of the functional connectivity weight between ROI i and the cortical network s, and N_M is the total number of cortical networks.

These PC values ranged from 0 to 1, with a greater value indicating grater connectivity (i.e., greater "hub"-like properties). Plots were thresholded with a voxel-wise value of 0.65 in order to visualize regions with connector hub-like properties.

References

[1] 1. Rangel BO, Hazeltine E, Wessel JR. Lingering Neural Representations of Past Task Features Adversely Affect Future Behavior. J Neurosci. 2023;43(2):282-92.

2. Cellier D, Petersen IT, Hwang K. Dynamics of Hierarchical Task Representations. J Neurosci. 2022;42(38):7276-84.

3. Gratton C, Sun H, Petersen SE. Control networks and hubs. Psychophysiology. 2018;55(3).

4. Power JD, Schlaggar BL, Lessov-Schlaggar CN, Petersen SE. Evidence for hubs in human functional brain networks. Neuron. 2013;79(4):798-813.

5. van den Heuvel MP, Sporns O. Network hubs in the human brain. Trends Cogn Sci. 2013;17(12):683-96.

6. Sporns O, Betzel RF. Modular Brain Networks. Annu Rev Psychol. 2016;67:613-40.

7. Bertolero MA, Yeo BT, D'Esposito M. The modular and integrative functional architecture of the human brain. PNAS. 2015.

8. Moneta N, Grossman S, Schuck NW. Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond. Trends Neurosci. 2024;47(12):1055-69.

9. Nassar MR, McGuire JT, Ritz H, Kable JW. Dissociable Forms of Uncertainty-Driven Representational Change Across the Human Brain. J Neurosci. 2019;39(9):1688-98.

10. Schuck NW, Cai MB, Wilson RC, Niv Y. Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. Neuron. 2016;91(6):1402-12.

11. Corbetta M, Akbudak E, Conturo TE, Synder AZ, Ollinger JM, Drury HA, et al. A common network of functional areas for attention and eye movements. Neuron. 1998;21:761-73.

12. Cole MW, Reynolds JR, Power JD, Repovs G, Anticevic A, Braver TS. Multi-task connectivity reveals flexible hubs for adaptive task control. Nat Neurosci. 2013;16(9):1348-55.

13. Rissman J, Gazzaley A, D'Esposito M. Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage. 2004;23(2):752-63.

14. Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb Cortex. 2018;28(9):3095-114.

15. Reber J, Hwang K, Bowren M, Bruss J, Mukherjee P, Tranel D, et al. Cognitive impairment after focal brain lesions is better predicted by damage to structural than functional network hubs. Proc Natl Acad Sci U S A. 2021;118(19).

16. Botvinick M, Braver TS, Carter CS, Cohen JD. Conflict monitoring and cognitive control. Psychological Review. 2001;108.

17. Freund MC, Bugg JM, Braver TS. A representational similarity analysis of cognitive control during color-word stroop. bioRxiv. 2021.

18. Jiang J, Wagner AD, Egner T. Integrated externally and internally generated task predictions jointly guide cognitive control in prefrontal cortex. Elife. 2018;7.

19. Rutledge RB, Dean M, Caplin A, Glimcher PW. Testing the reward prediction error hypothesis with an axiomatic model. J Neurosci. 2010;30(40):13525-36.

20. Konishi M, McLaren DG, Engen H, Smallwood J. Shaped by the Past: The Default Mode Network Supports Cognition that Is Independent of Immediate Perceptual Input. PLoS One. 2015;10(6):e0132209.

21. Vatansever D, Menon DK, Manktelow AE, Sahakian BJ, Stamatakis EA. Default Mode Dynamics for Global Functional Integration. J Neurosci. 2015;35(46):15254-62.

22. Bertolero MA, Yeo BTT, Bassett DS, D'Esposito M. A mechanistic model of connector hubs, modularity and cognition. Nat Hum Behav. 2018;2(10):765-77.

23. Cocuzza CV, Ito T, Schultz D, Bassett DS, Cole MW. Flexible Coordinator and Switcher Hubs for Adaptive Task Control. J Neurosci. 2020;40(36):6949-68.
24. Egner T, Hirsch J. Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. Nat Neurosci. 2005;8(12):1784-90.
25. Miller BT, D'Esposito M. Searching for "the top" in top-down control. Neuron. 2005;48(4):535-8.
26. Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nat Neurosci. 2006;9(11):1432-8.
27. van Bergen RS, Ma WJ, Pratte MS, Jehee JF. Sensory uncertainty decoded from visual cortex predicts behavior. Nat Neurosci. 2015;18(12):1728-30.
28. Vul E, Hanus D, Kanwisher N. Attention as inference: selection is probabilistic; responses are all-or-none samples. J Exp Psychol Gen. 2009;138(4):546-60.
29. Li H, Sprague TC, Yoo AH, Ma WJ, Curtis CE. Neural mechanisms of resource allocation in working memory. 2024.
30. Behrens TE, Woolrich MW, Walton ME, Rushworth MF. Learning the value of information in an uncertain world. Nat Neurosci. 2007;10(9):1214-21.
31. Jiang J, Heller K, Egner T. Bayesian modeling of flexible cognitive control. Neurosci Biobehav Rev. 2014;46 Pt 1:30-43.
32. Esteban O, Markiewicz CJ, Blair RW, Moodie CA, Isik AI, Erramuzpe A, et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods. 2019;16(1):111-6.
33. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage. 2011;54.
34. Behzadi Y, Restom K, Liau J, Liu TT. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. Neuroimage. 2007;37(1):90-101.
35. Cox RW. AFNI Software for analysis and visualization of functional magnetic resonance neuroimages. 1996.
36. Pedregosa F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. . Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
37. McLaren DG, Ries ML, Xu G, Johnson SC. A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. Neuroimage. 2012;61(4):1277-86.
38. Gratton C, Nomura EM, Perez F, D'Esposito M. Focal brain lesions to critical locations cause widespread disruption of the modular organization of the brain. Journal of Cognitive Neuroscience. 2012;24.
39. Hwang K, Shine JM, Bruss J, Tranel D, Boes A. Neuropsychological evidence of multi-domain network hubs in the human thalamus. Elife. 2021;10.

Acknowledgments

Author contributions are as follows.

Conceptualization: SCL, KH, JJ

Methodology: SCL, KH, JJ

Investigation: SCL, HH, KH, JJ

Visualization: SCL, KH, JJ

Supervision: KH, JJ

Writing—original draft: SCL, KH, JJ

Writing—review & editing: SCL, KH, JJ

Authors declare that they have no competing interest.

All data and code will be made available upon publication.