Title

How Task Representations Integrate Information from Multiple Sources

Authors

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

* kai-hwang@uiowa.edu

Affiliations

[1] The University of Iowa, Iowa City, IA.

[2] Cognitive Control Collaborative (CCC).

† These authors contributed equally to this work.

Abstract

Adaptive behavior requires integrating multiple sources of information, yet neural and computational mechanisms underlying this process are poorly understood. We introduce a novel Bayesian computational framework elucidating how internally maintained and externally cued information are integrated to guide goal-directed actions. Our model infers trial-by-trial beliefs for inputs and integrates them into a joint probability distribution, where entropy quantifies overall uncertainty. This integrated representation predicts a probabilistic task belief and generates continuous measures of prediction errors and updating. Decoding analyses revealed distinct neural substrates encoded external and internal inputs, with convergence in frontoparietal hub regions where activity fluctuations scaled with entropy. Increasing integrated uncertainty modulated connectivity patterns, biasing the generation of integrated task outputs and prediction errors. Together, these findings capture the full integration process—encoding inputs, forming an integrated representation jointly encoding uncertainty, and producing a unified task output. This work provides novel insights into neural mechanisms underlying cognitive integration.

Teaser

How frontoparietal hubs integrate internal state and sensory inputs, resolving uncertainty to guide adaptive behavior.

Introduction

The ability to engage in complex goal-directed behavior is a hallmark of human intelligence and essential for navigating everyday life. To achieve goals, humans construct rich cognitive representations that capture rule structures and contextual contingencies (Cellier, Petersen, & Hwang, 2022; Rangel, Hazeltine, & Wessel, 2023). These cognitive representations transform sensory and mnemonic inputs into actions and decisions to guide goal-directed behavior. Critically, this goal directed behavior often involves integrating multiple sources of information. For example, when driving, you might see blue and red flashing lights (an external perceptual cue) in the distance due to a car accident. To get home as quickly as possible, this perceptual information must be combined with internally tracked and maintained expectations or beliefs, such as the

anticipated traffic congestion or road closures on alternative routes home. However, the neural and cognitive processes supporting this adaptive behavior remain unclear. Therefore, "cognitive integration" – the process of integrating abstracted contextual information (i.e., not directly observable nor mapped to an external sensory cue) with externally presented perceptual inputs to guide behavior – is a fundamental challenge to human neurocognitive systems. In this paper, we propose a novel computational framework that delineates the mechanisms by which internal representations and external inputs are integrated to guide adaptive behavior.

Existing research investigating cognitive integration can be loosely categorized into two lines of inquiry: (1) functional brain network hubs and (2) hierarchical control. Functional brain network hubs are domain-general regions with extensive connectivity across multiple systems (Gratton, Sun, & Petersen, 2018; Power, Schlaggar, Lessov-Schlaggar, & Petersen, 2013; van den Heuvel & Sporns, 2013). It has been proposed that placing integrative functions in these hub-like regions facilitates broad access to information across distributed systems (Sporns & Betzel, 2016). By leveraging this connectivity architecture, hubs can flexibly receive inputs and transmit outputs among diverse brain networks, thereby supporting integration. Studies of functional brain networks often identify regions in the frontoparietal network (FPN) and the default mode network (DMN) as exhibiting strong hub properties (Bertolero, Yeo, & D'Esposito, 2015). These regions also display dynamic, task-evoked connectivity with distributed systems (Bertolero, Yeo, Bassett, & D'Esposito, 2018; Cole et al., 2013; Vatansever, Menon, Manktelow, Sahakian, & Stamatakis, 2015). Although these findings suggest that hub-like brain regions are architecturally positioned to support integration, connectivity patterns alone do not explain the underlying neural and cognitive mechanisms of integration. Studies focusing on characterizing connectivity patterns do not test the computational processes that enable integration within hubs, nor do they address the representational structure of the cognitive representations encoded in these putative hubs. As such, it remains unclear how exactly hub regions integrate information from diverse input sources.

In contrast, the hierarchical control framework has focused on investigating the representational structure of cognitive representations. Research has shown that more concrete, sensory-motor cognitive representations are encoded in more caudal region of the prefrontal cortex, whereas more abstract representations are encoded more rostrally (Badre & D'Esposito, 2007; Badre, Hoffman, Cooney, & D'Esposito, 2009; Collins, Cavanagh, & Frank, 2014; Nee & D'Esposito, 2016). In this framework, cortico-subcortical loops propagate information from rostrally encoded abstractions to progressively more caudal, concrete regions, ultimately driving motor outputs. However, it is unclear how the hierarchical framework accounts for integration when input information sources include latent, internally maintained cognitive representations that lack direct external cues. Moreover, it does not address how uncertain, or ambiguous, inputs are integrated. More importantly, although this hierarchical model provides depiction of top-down, abstract-to-concrete information flow, it does not address how multiple, diverse sources of information are integrated.

Our goal is to build a framework that specifies how internally maintained and externally cued information is integrated. To achieve this, the framework must clearly define several elements. First, it should specify the structure of the inputs to be integrated (i.e., precisely what information is entering the system). Second, it needs to define the processes that transform these inputs into an integrated representation, including the

representational format of that integrated representation. Third, it should characterize the outputs that can be extracted from this integrated representation. Finally, it should specify the updating mechanisms that modify the inputs as needed based on information from both the output(s) and integrated product. Critically, computational models that incorporate these elements can guide neuroimaging to investigate the associated neural processes. When combined with neuroimaging techniques, this model can provide trial-by-trial predictors to probe neuronal underpinnings in a theoretically guided way, thereby increasing the precision of brain-behavioral investigations.

In the present study, we apply this approach to determine how information from cognitive states (i.e., internal beliefs or expectations) are integrated with perceptual inputs to guide actions and decisions. The inclusion of a latent cognitive state belief is significant because existing work has largely focused on how external sensory cues are linked to task rules or context (Cellier et al., 2022; Cole, Bassett, Power, Braver, & Petersen, 2014; Cole et al., 2013; Hwang, Shine, & D'Esposito, 2019). Yet, contextual information is often not directly observable from the environment. Consider again the driving example: if we decide to take an alternative route to avoid a traffic jam, we do so based on internally maintained knowledge about typical congestion patterns at this time of day—knowledge not tied to any immediate perceptual cue. This illustrates how a latent cognitive state, independent of current sensory inputs, allows you to anticipate and plan ahead; thereby underscoring the importance of integrating both internal and external information for goal-directed behavior.

We developed a computational model based on a Bayesian framework to infer latent variables that represent subjective beliefs from (1) externally presented, (2) internally maintained information sources (inputs), and (3) the readout of the integrated representation (output). To mechanistically define how these inputs are transformed into an integrated representation, the model posits that internal and external sources of information are integrated into a joint distribution, which serves as an integrated cognitive representation. This joint representation maps onto a conditional probabilistic distribution, to predict the output of integration given the specific inputs. As such, the probability of performing a given action or decision is conditioned on the joint contribution of both external and internal sources (Fig. **1A**).

The latent variables in our model estimate not only the categorical, dichotomous belief of each input and output, but also the corresponding levels of uncertainty in these beliefs. For example, the latent variable estimating the internally maintained cognitive state indicates both the participant's current dichotomous state belief and their continuous uncertainty in that belief. The combined uncertainties of the multiple input sources that contribute to this joint probability distribution are reflected in properties, such as entropy, of this joint distribution. Thus, entropy can serve as a measure of the uncertainty associated with the integrated representation (Fig. **1B**). If both sources of information show low uncertainty, then the estimated entropy of the joint distribution would be low, meaning the degree of integration is high (Fig. **1B** left). Conversely, if both to-be-integrated sources show greater uncertainty, then the estimated entropy will be higher, meaning a lower degree of integration (Fig. **1B** right). Intermediate levels of uncertainty produce entropy values that fall between these extremes, thereby illustrating how the model can capture varying degrees of integration.

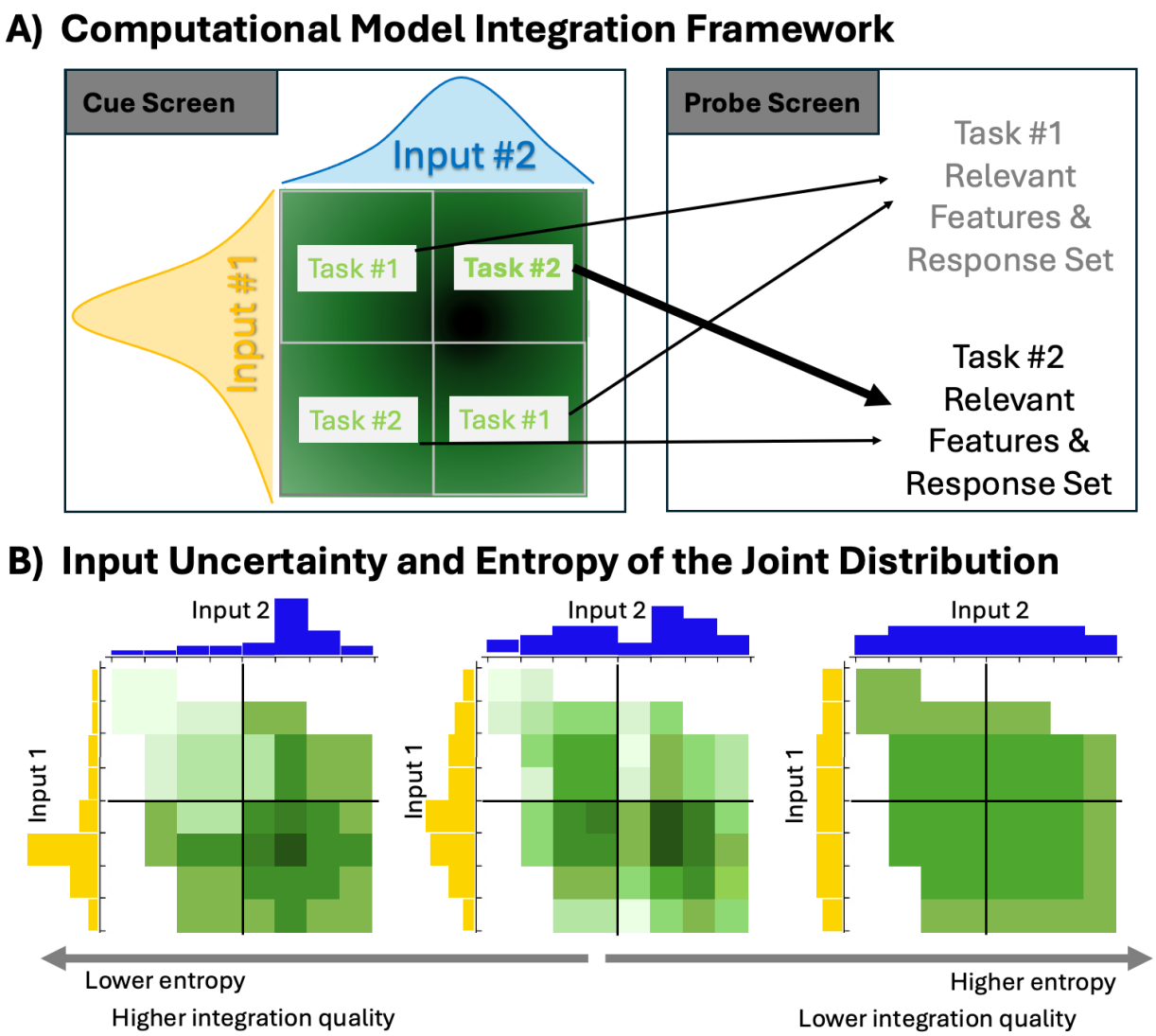

Figure 1. (**A**) Depiction of integration of information into a joint distribution and its corresponding conditional probability distribution that predicts the specific task or decision to perform. (**B**) Depiction of entropy estimate of the joint distribution of two input variables: state and color. Low entropy resulting from low uncertainty in both inputs. Intermediate entropy due to moderate uncertainty in one or both inputs, whereas high entropy is due to very high uncertainty in the inputs.

Furthermore, because the model estimates continuous measures of uncertainty, we can leverage this information to investigate updating processes and the signals that trigger them by calculating continuous measures of the magnitude of the change in the belief and the prediction errors, separately for each model belief. For example, if a participant is highly confident in their state belief, an incorrect feedback event should produce a large prediction error; conversely, if this participant is highly uncertain about their state belief, the same incorrect feedback should yield a relatively smaller prediction error. Likewise, when prediction errors are larger, updating processes need to induce a greater degree of change in that input belief.

In summary, our framework provides trial-by-trial predictors that capture the structure and uncertainty of input and output representations, the quality of integration via the joint probability distribution, its entropy, and prediction errors signaling the need for

belief updating as well as updating processes. We then combined these model-derived estimates with fMRI using parametric modulation and decoding analyses to investigate the neural mechanisms underlying cognitive integration. Finally, we conducted functional connectivity analyses to test the influence of the inputs on the integrated product and the influence of the integrated product on the output.

Results

To address cognitive integration, we designed a task paradigm in which participants integrated one external and one internal source of 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**).

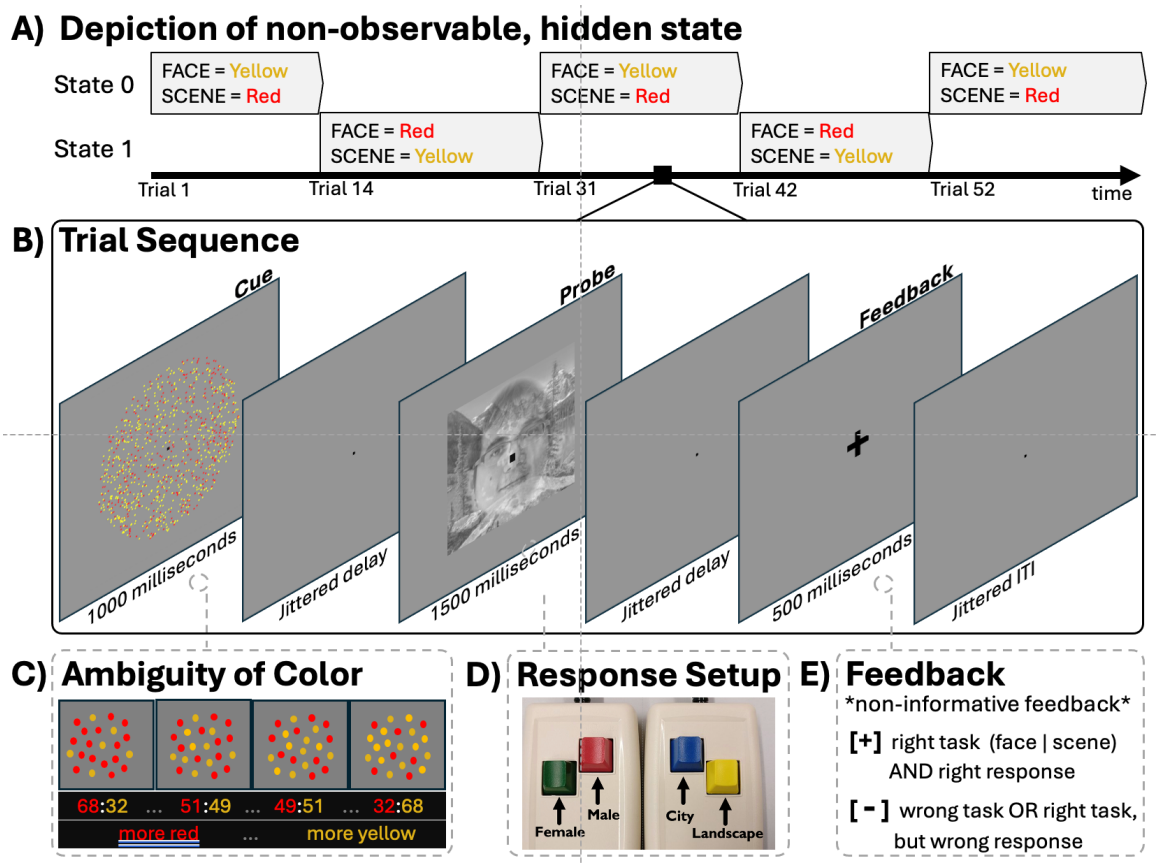

Figure 2. (**A**) Depiction of non-observable state that changes every 10-20 trials. **(B)** Trial sequence example: cue screen (1000ms), jittered delay (range: 0.5-10.5 seconds, average=3.75 seconds), probe screen (1500ms), jittered delay (range: 0.5-10.5 seconds, average=3.75 seconds), feedback screen (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 task. This model estimates two probabilistic distributions—the volatile hidden state and the ambiguous perceptual cue, or color ratio—as inputs (i.e., sources of information) to be integrated. This model was applied separately to each subject's behavioral data (see one example subject in Fig. **3B**). By varying the ambiguity, or uncertainty, of these inputs across trials we manipulated integration quality. The joint probability of these inputs provided a direct measure of integration while the entropy of the joint distribution provided an indirect measure of integration quality. This overall entropy influenced the model output (i.e., the task to be performed). By tracking uncertainty within both inputs and outputs, we could investigate how these representations are updated on a trial-by-trial basis using prediction errors and the magnitude of change in state belief (i.e., state derivative). In short. this model outlined a mechanistic framework of integration by constructing a joint distribution of two inputs, one observable and one non-observable. The model then provided specific predictors that captured the trial-by-trial variation of inputs, integration (e.g., the entropy of the joint distribution), and the integrated outputs. Moreover, this model avoided defaulting to a homunculus explanation of control by providing specific predictors that captured updating processes and the prediction errors signaling the need to engage these processes. These measures allowed us to link integrative processes and the resulting joint representation to behavioral and neuroimaging data with a theoretically constrained approach.

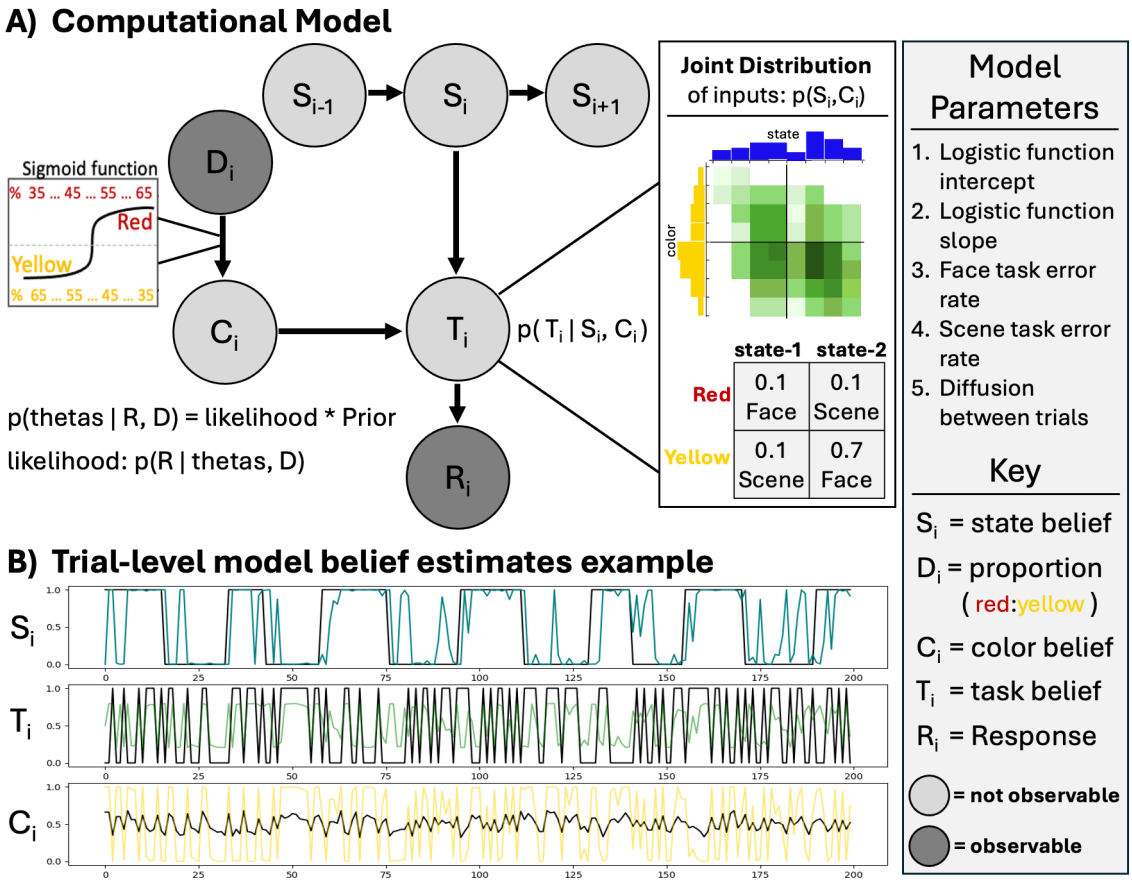

Figure 3. (**A**) Computational (Bayesian generative) model diagram. Observable information (R_i and D_i) used to infer three latent variables (S_i, C_i, and T_i). (**B**) Trial-level model estimated "beliefs" from one example subject. Black lines: true state, task, or color. Blue, green, and yellow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

The key feature of our model is that both sources of information are encoded probabilistically, enabling their uncertainty to be integrated into a joint representation in order to determine the correct task based on these inputs. Given that the key assumption of this model is probabilistic integration of inputs, we compared it to four alternative models without such assumption: (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 model 2, except that for the most ambiguous color proportions, participants guess randomly, and consequently do not update their state belief after feedback (pSjP). This model was included to capture the possibility that the participant used two different strategies (i.e., random guessing for high color uncertainty and encoding a dichotomous color belief using the softmax function for low color uncertainty).

In the alternative models, treating state or color as dichotomous means that subject beliefs do not include uncertainty. Instead, a dichotomous belief assumes that only the final category decision (e.g., red or yellow; state 1 or state 2) is represented. In such cases, integration would proceed in a stepwise decision process, without requiring the formation of a probabilistic joint representation. Therefore, these alternative models do not assume integration through a joint probabilistic representation. Consequently, belief updating is also treated as an all-or-none process - expectations are either fully violated or not and beliefs are 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.

The proposed model and alternative models were fit to the behavioral data from the 38 usable subjects who completed the fMRI task paradigm. In terms of prediction error (Fig. **4A**), the proposed computational model (both state and color probabilistic; pSpP prediction error rate (PE)=23.08%) outperformed the three alternative models (pSdP PE=25.03%, dSpP PE=33.70%, dSdP PE=24.93%, and pSjP PE=50.04%). Protected exceedance probability analysis on Bayesian information criterion (BIC; Fig. **4B**) showed strong preference of the proposed integration model (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 perception in the subject-level data and the fact that our model assuming probabilistic color outperforms the model assuming random guessing at very ambiguous ranges. Overall, these results provide empirical support for the key assumptions of our integrative model.

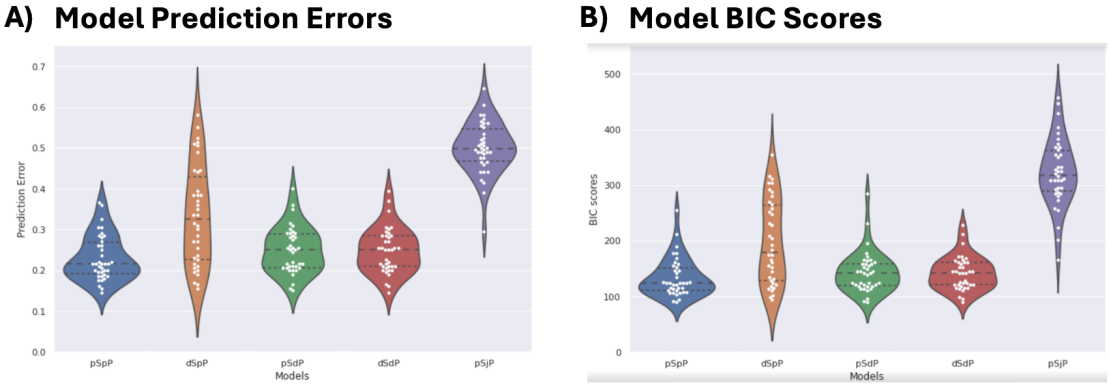

Figure 4. (**A**) Prediction error (percent of trials) of each computational model. Each dot represents the prediction accuracy for an individual subject. (**B**) BIC scores for each computational model. Each dot represents the BIC score for an individual subject.

2.2 Integration Uncertainty and Response Times: If the entropy of the joint distribution is a measure of the degree of integration, then we would expect slower response times with greater entropy. To test this, participant response times were entered as the dependent variable in a mixed effects regression with the winsorized entropy values as fixed effects independent variables. Participants were entered as random intercepts. The entropy values were not included in the random effects structure because including them as a random slope did not significantly improve model fit, chi(2)=4.918, p=.086. As expected, greater entropy of the joint distribution predicted slower response times (β= 0.028, CI = 0.01 to 0.05, t(7572)=3.064, p=.002). We calculated the effect size using Cohen's d and found a medium effect size, d=0.50,

Given that the task belief is derived from the joint distribution, we would also expect greater task uncertainty to be associated with slower response times. To test this,

we used a mixed-effects regression model to examine the relationship between response times and task uncertainty (Ti), the latent variable reflecting the task-to-perform. In the regression model, task uncertainty was included as a fixed effect, with random intercepts for participants. Task uncertainty was not included as a random slope because its inclusion did not significantly improve model fit, chi(2)=5.945, p=.051. As expected, the uncertainty of the task-to-perform was associated with slower response times (β= 0.038, t(7547)=2.257, p=.024, CI: 0.005 to 0.71). We calculated the effect size using Cohen's d and found a medium effect size, d=0.37. Together, these findings indicated that both entropy of the joint distribution and task uncertainty were associated with slower response times, thereby supporting the behavioral relevance of our computational model.

2.3 Anatomical Underpinnings of Integration: Our model-driven approach enabled us to investigate the neural substrates associated with each component of integration process—inputs, joint representation, output, and prediction errors and updating. This framework sheds light on two critical aspects: (1) how information is transformed from individual sources into a joint representation and the related integration uncertainty (i.e., entropy) that biases the task (output) belief and (2) how trial-level feedback is used to update information at the input level.

From our computational model, we conducted the following whole-brain GLM analyses: First, to determine the neural representations of integration inputs (state and color) and output (task), we performed decoding analyses to predict trial-by-trial beliefs in state (Si), color (Ci), and task (Ti). Second, to identify neural substrates associated with input and output uncertainty, we applied parametric modulation analyses using the model-derived uncertainty estimates for each belief. Third, to examine the neural substrates that implement integration itself, we performed decoding analyses to predict the trial-by-trial joint probability, which is the specific combination of color and state. Additionally, we included the entropy of the joint representation as a parametric regressor to assess the overall uncertainty in this integrated representation. Fourth, to identify neural substrates associated with updating internally tracked information, such as state belief, we applied parametric modulation analyses using the model-derived prediction errors and derivative estimates for the state belief. Finally, we performed connectivity analyses to assess task-related connectivity patterns between ROIs during the cue and probe epochs, and to assess how trial-by-trial entropy values modulate these connectivity patterns.

2.3.1 Inputs for Integration: The inputs for integration were an (1) externally presented color cue and (2) internally maintained and continuously tracked state. The state required continuous monitoring and updating based on trial-level feedback, as no external indicators explicitly signaled the state. In contrast, the color cue, which changed on a trial-by-trial basis, was updated/encoded when visually presented at the start of each trial. Due to these differences, we expected that the neural substrates encoding these two sources would be dissociable.

2.3.1.1 State Belief Decoding: Prior research suggests the orbitofrontal cortex (OFC) encodes state information (Moneta, Grossman, & Schuck, 2024; Nassar, McGuire, Ritz, & Kable, 2019; Schuck, Cai, Wilson, & Niv, 2016). Based on this, we hypothesized that OFC regions would show a significant correlation between the trial-level state beliefs predicted by our computational model and those estimated by the decoding analysis (Fig. **5A**).

Following cue onset, we found significant clusters in the parahippocampus, fusiform gyrus, middle frontal gyrus (MFG), and visual association cortex (Fig. **5B**). Following probe onset, we once again found a significant cluster in the visual association cortex in addition to significant clusters in the middle temporal gyrus (MTG), primary sensory cortex, and superior parietal lobe (SPL). Contrary to our prediction, we did not find a significant cluster in OFC when using our a priori defined minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005). Because of this, we did an exploratory follow-up and found that a voxel-level t-value threshold of 2.412 does show a significant cluster (21 voxels) in OFC.

2.3.1.2 Color Belief Decoding: Previous research has shown that frontoparietal network activity increases when participants process externally presented cues linked to rules guiding behavior (Cole 2017). Additionally, such cues often engage regions within both the dorsal and ventral attention networks (Corbetta et al., 1998). 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 (the color of the cue informs the subsequent task-to-perform), we hypothesized a correlation between model-estimated trial-level color beliefs and decoding-predicted trial-level color beliefs in 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 hypothesis, following cue onset our decoding analysis revealed distributed clusters across the prefrontal cortex, temporal cortex, limbic, occipital, and parietal cortices (Fig. **5C**). This included significant clusters within the ventral attention networks, such as MT and IPS. We also found a significant cluster lateral to the fusiform gyrus. Additionally, we found clusters in regions typically associated with the frontoparietal network, such as the right and left superior frontal gyrus (SFG), middle frontal gyrus (MFG), and superior parietal lobule (SPL). Unexpectedly we also found a significant cluster in OFC with color decoding. Following probe onset, we found significant clusters in inferior frontal cortex, such as OFC and IFG, superior frontal cortex, such as MFG, preSMA, and primary motor cortex, as well as inferior temporal cortex, such as fusiform gyrus.

These findings suggest that regions typically associated with both the visual attention and frontoparietal networks encode information about externally presented cues that are tied to task rules. However, the question remains as to how these regions integrate externally presented information with internally maintained information, especially since state and color belief is decodable only from distinct regions in distinct networks.

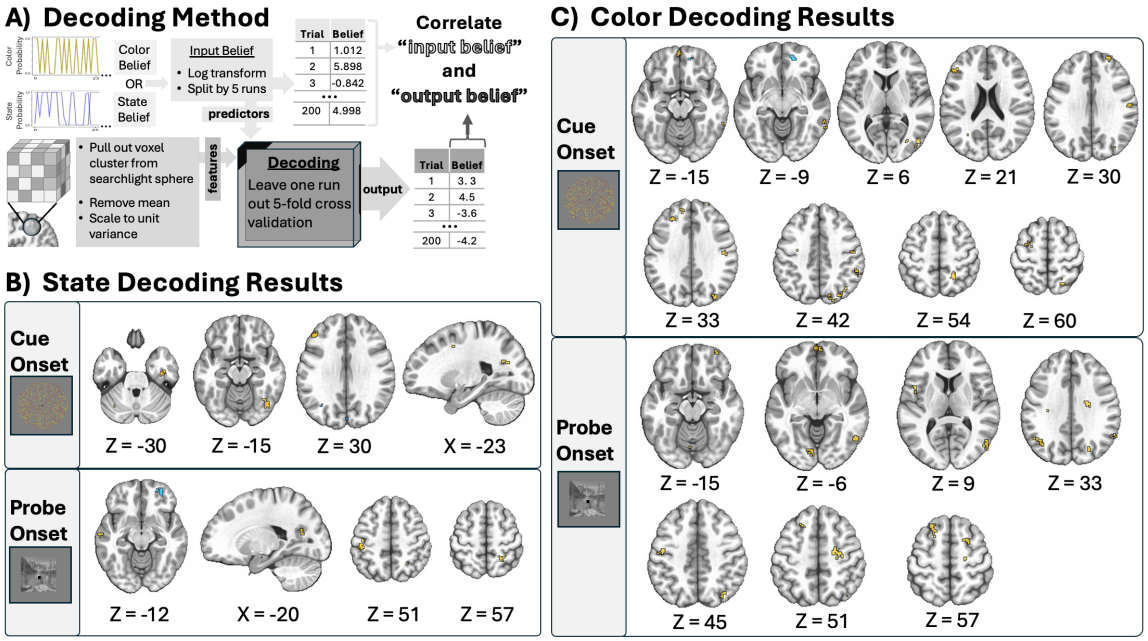

Figure 5. (**A**) Depiction of decoding method for color and state belief. Decoding results for color belief (**C**) and state (**B**) belief with cue and probe epochs. Results based on minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005).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

2.3.1.3 State and Color Uncertainty Parametric Modulation: One key prediction is that regions encoding/tracking inputs for integration would modulate their activity profiles based on the level of uncertainty in these inputs. Specifically, regions tracking state and/or color beliefs were expected to exhibit greater activity in response to higher state or color uncertainty (Fig. **6A**). Most likely because greater uncertainty demands increased effort to either resolve the ambiguity or encode and maintain the uncertainty of the state or color belief. We therefore first determine the neural substrates that encode state and color belief uncertainties.

We observed substantial overlap in brain regions exhibiting significant activity in response to greater uncertainty in both color (Fig. **6B**) and state (Fig. **6C**) beliefs. Specifically, MFG, medial prefrontal cortex (mPFC), and IPS, showed this overlap, as well as ventrolateral prefrontal cortex (vlPFC). A cluster in the inferior frontal sulcus (IFS) overlapped between state decoding and parametric modulation with state uncertainty. Other regions where only state information was decoded did not exhibit state or color uncertainty parametric modulation. A cluster in the fusiform gyrus overlapped between color decoding and parametric modulation with color uncertainty. Other regions where only color information was decoded did not show color or state uncertainty parametric modulation.

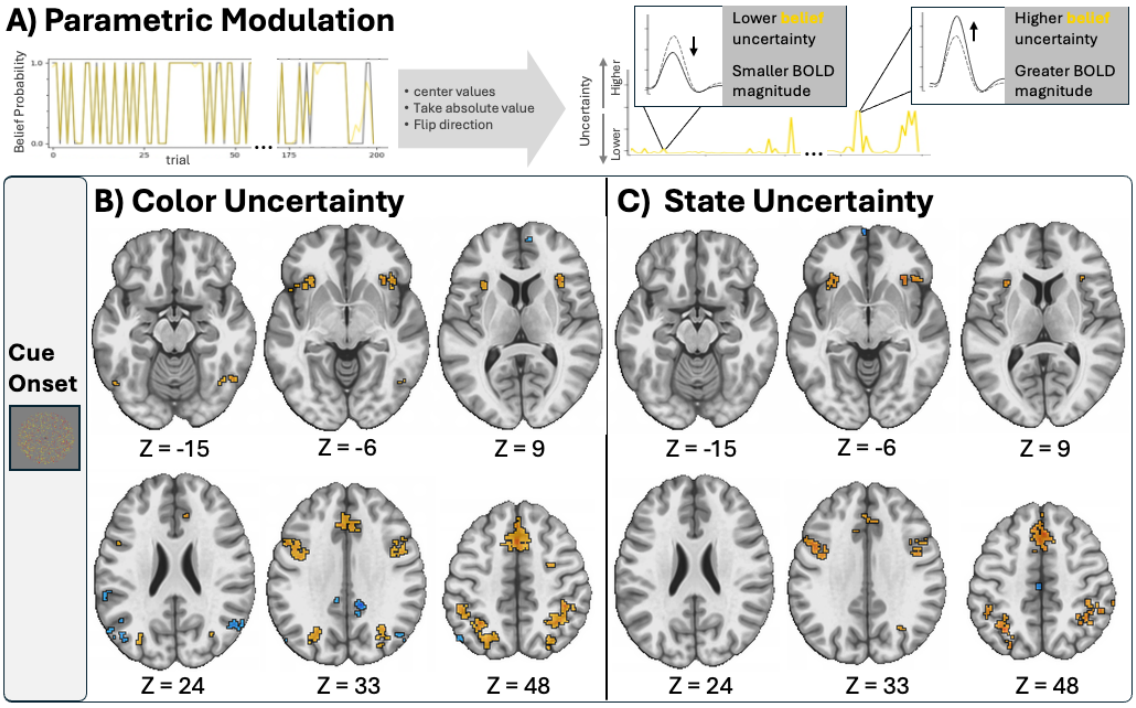

Figure 6. (**A**) Depiction of parametric modulation analyses for color uncertainty and state uncertainty. Uncertainty calculated by centering color/state belief, taking the absolute value, and flipping the direction so that a larger value reflects greater uncertainty. Color (**B**) and state (**C**) 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).

2.3.2 Joint Representation and the Output from Integration: Based on our framework of cognitive integration, the joint representation can be operationalized as a posterior distribution of multiple probabilistic inputs. Here, the joint probability, which is the specific combination of color and state, can be used to predict the output of integration. The output of the joint representation is the task belief, which specifies the task-to-perform based on integrated color and state beliefs. Following integration of this information at cue onset, we expect 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, and regions involved in maintaining task rules long-term, such as hippocampal regions as well as fusiform face area (FFA) and parahippocampal place area (PPA) regions given that the task-to-perform was making a discrimination judgement about either a face or scene image.

2.3.2.1 Joint Probability and Task Belief Decoding: The joint probability reflected the specific combination of color and state on each trial. In our task paradigm, there were four possible combinations of color and state on any given trial. Given that the specific combination of state and color directly informs the task to perform, we expected to decode both the joint probability and task belief from hub regions in the frontoparietal network, such as MFG and IPS, given that these regions are optimally positioned to receive and integrate distinct sources of information and then bias downstream task regions to control behavior. Moreover, we also expected to decode the joint probability and task (output) belief from task specific regions such as hippocampal regions and FFA/PPA.

In line with these expectations, we found significant clusters in MFG and IPS as well as hippocampal regions and FFA/PPA (Fig. 7A). There were also robust clusters spanning from supplementary motor cortex regions across the primary motor and sensory cortices towards the IPS. These same clusters were found when applying decoding analyses at probe onset.

Correlating decoding strength at cue onset with model estimated participant task belief revealed distributed clusters in many of the same regions where the joint probability was decoded (Fig. 7B). Of particular note, task belief was decoded from the same regions that spanned from supplementary motor cortex regions across the primary motor and sensory cortices into the IPS. Moreover, we found these same clusters when applying decoding analyses at probe onset. This pattern may reflect that participants could preparatorily activate the left or right hand depending on which task they believed they should perform.

2.3.2.2 Joint Representation and Task Uncertainty: We further conducted a parametric modulation analysis to assess which regions fluctuate with the combined overall uncertainty (entropy) and output (task) uncertainty. Entropy quantifies the summed uncertainty across all possible outcomes, reflecting the overall uncertainty that is not specific to any single input. In contrast, the joint probability reflected the specific combination of inputs. Entropy remains low when both inputs have low uncertainty, high when both are highly uncertain, and intermediate when their uncertainty levels differ. Thus, by calculating the distribution's entropy in our computational model, we obtain a proxy for the combined uncertainty of the joint representation.

Consequently, if any regions fluctuated their activity profiles with both state and color uncertainty, we expected these same regions to also fluctuate their activity profiles with entropy estimates, given entropy's role as a measure of combined uncertainty. Moreover, since the joint probability and task belief were decoded from largely overlapping regions, we expected these same regions to fluctuate with task uncertainty. Consistent with these expectations, following cue onset, we found that regions whose activity scaled with higher entropy or task uncertainty significantly overlapped with those responding to increased uncertainty in state and color beliefs. More specifically, MFG, mPFC, IPS, SFG, and vlPFC showed this overlap (Fig. **7C-D**).

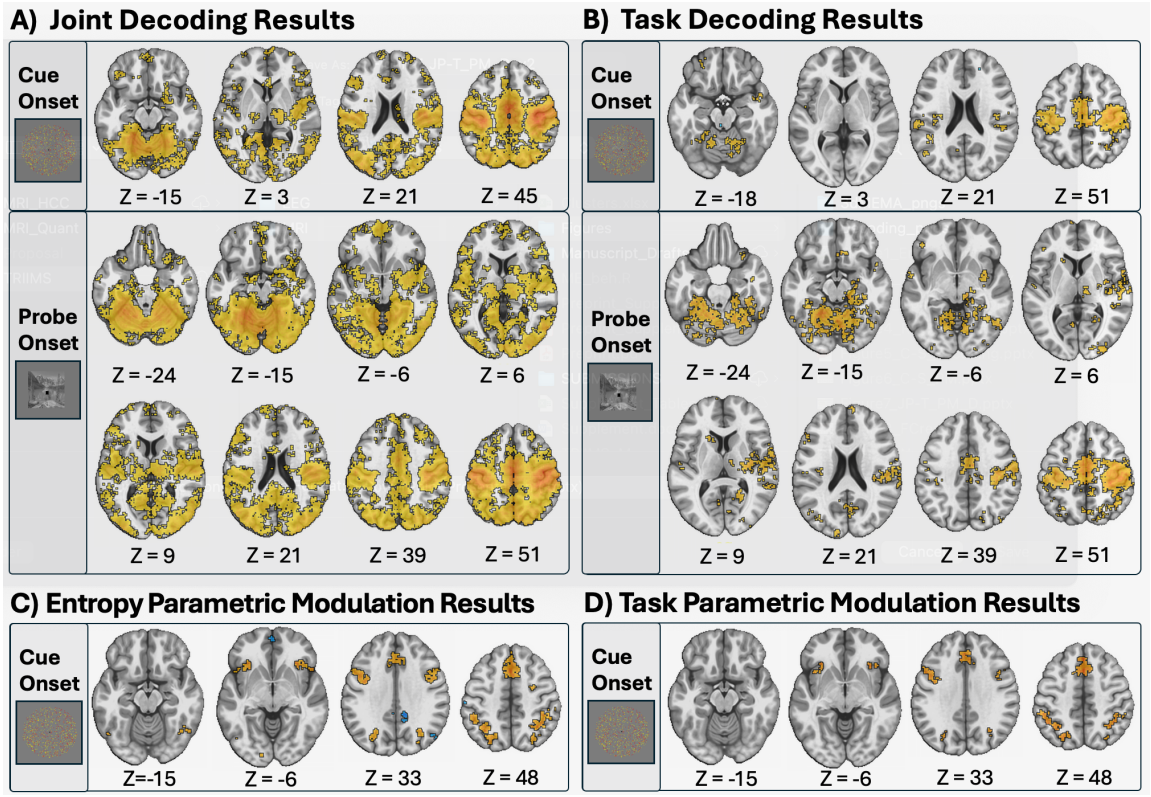

Figure 7. Decoding results (**A-B**) for joint probability (**A**) and task belief (**B**) with cue and probe epochs based on minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005). Parametric modulation results (**C-D**) for entropy (**C**) 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.

In sum, regions such as MFG and IPS, have thus far shown a couple key findings. First, they fluctuate their activity patterns with input uncertainty, entropy, and output uncertainty. Second, activity patterns in these regions can be used to decode the joint probability as well as the output (task) belief. 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. Next, we performed a direct test of functional connectivity patterns between regions that encode input sources, entropy, and task outputs to test this interpretation.

2.3.3 Functional Connectivity: So far, our results indicate that frontoparietal regions—known for their flexible and diverse connectivity across various cognitive control tasks (Cole et al., 2014; Cole, Ito, Cocuzza, & Sanchez-Romero, 2021; Cole et al., 2013)—are modulated by entropy in our task. This finding suggests that these regions may detect the combined uncertainty of task-relevant inputs and convert this estimate into a biasing signal that influences task output. To further test this interpretation, we conducted a novel task-related connectivity analysis.

Briefly, we conducted beta-series connectivity (Rissman et al., 2004) analyses using single-trial beta estimates from all trials across different epochs (cue, probe, feedback). This connectivity analysis was conducted between 400 cortical ROIs (Schaefer

et al., 2018). 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 the entropy measure included as a moderator. This approach yielded estimates for both the main effect of connectivity and its interaction with the integrated uncertainty. For the purpose of this analysis, we focused on the interaction effect. 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 the target ROI changes as the entropy of the joint representation increases. We predicted that during the cue epoch, with higher uncertainty, the coupling between regions that encode entropy and task inputs (color and state) may strengthen. This would suggest 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 encoding the task output and entropy. This would reflect networks dynamically adjusting communication to select and execute this selected task-to-perform.

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

We found that the cue and probe epochs exhibited distinct brain-wide connectivity patterns (Fig. **8A-B**). During the cue epoch, higher integrated uncertainty (entropy) was associated with a significant positive modulation between source entropy and target color, t=3.787, $p<.001$, and a significant negative modulation between source entropy and target task, t=-5.665, $p<.001$ (Fig. **8D-E**). This pattern suggests that, at cue onset, increased uncertainty enhances connectivity related to processing external sensory information while decreasing connectivity with regions that encode task output. In contrast, during the probe epoch, the connectivity pattern shifted: significant positive modulation emerged between source entropy and target task, t=4.176, p<.001 (Fig. **8B**), indicating that integrated uncertainty primarily influences task-related connectivity as decisions are being executed. These findings demonstrate that as integrated uncertainty increases, connectivity patterns are dynamically coupled. Specifically, during the probe epoch, increased connectivity with task output regions may reflect top-down modulation that biases network interactions to implement the selected task when all necessary information is available. To further test the directional effect on biasing task output, we used source entropy from cue epochs to predict target task from probe epochs (Fig. **8C**). We found that higher integrated uncertainty (entropy) was associated with a significant positive modulation between source entropy (from cue epochs) and target task (from probe epochs), t=6.220, $p<.001$ (Fig. **8F**). These results further support our interpretation that top-down influences from cue to probe epochs, with integrated uncertainty measured after cue onset shaping task-related output processing at probe onset when decisions are executed.

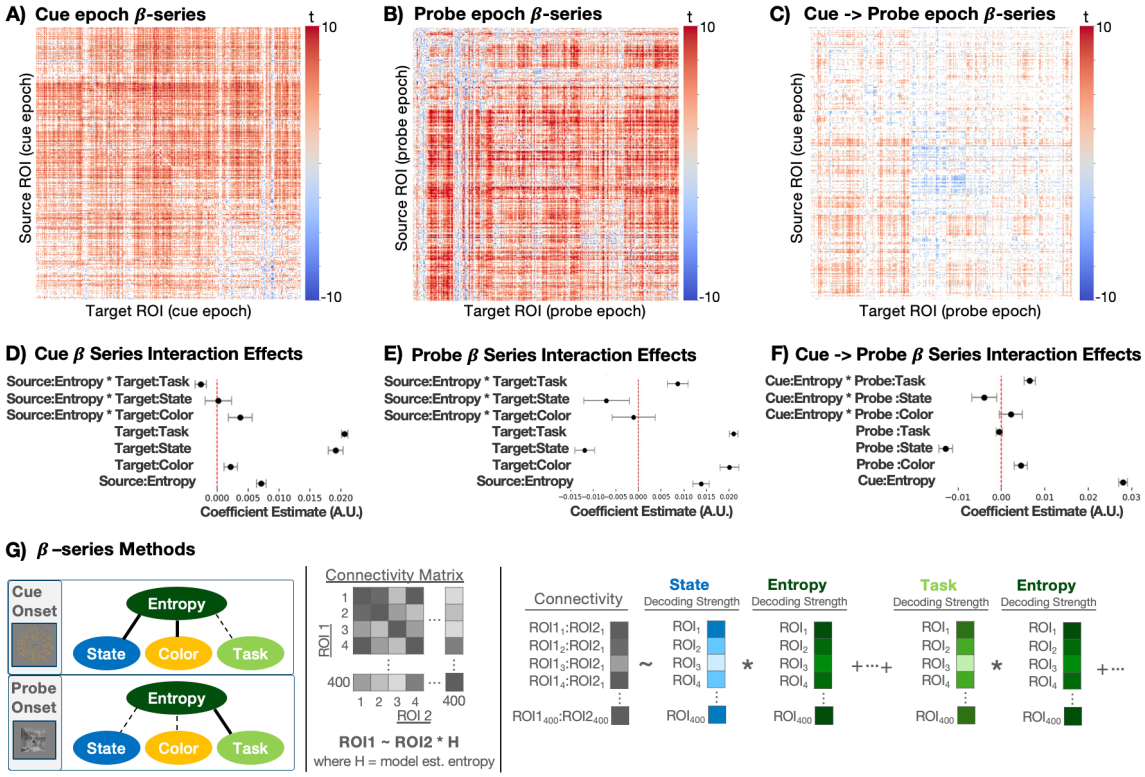

Figure 8. Beta-series connectivity matrices of entropy interaction for the cue (**A**) and probe (**B**) epochs as well as cue-to-probe (**C**) epochs, displaying T-values from a mixed-effects regression analysis. Each cell represents the effect of a "source" ROI (rows) on a "target" ROI (columns). Warmer 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. The face stimulus in the probe example is of the corresponding author.

2.3.4 Functional Hubs: Our integration framework predicts that regions responsible for integrating internal and external information—such as the MFG and IPS—should exhibit hub-like properties. Specifically, these regions are expected to dynamically track the combined uncertainty of task-relevant inputs and convert this uncertainty into a biasing signal that guides task output. To test this prediction, we computed voxel-wise participation coefficient (PC) scores using a previously developed method (Reber et al., 2021). PC estimates ranged from 0-1 and we a priori defined hub regions as regions with a

PC > 0.65. These results show that MFG and IPS are hub regions based on these criteria (Supplementary Figure 1). Moreover, we show that there is a strong overlap between regions fluctuating with entropy/uncertainty and hub regions. This further supports our account of integration where hub regions read the combined uncertainty of task-relevant inputs and convert this estimate into a bias signal that influences task output.

2.3.5 Updating Belief Representations: Previous sections have focused on the integration of information and how this joint representation biases the output, or task-to-perform. However, our account of integration would be incomplete without considering how the inputs to integration are updated following feedback. In our model, the state belief required continuous tracking and updating based on trial-level feedback, as no external indicators explicitly signaled the state on any given trial. By contrast, the color belief could be updated immediately following the cue onset. Therefore, state belief updating processes required the subjects to make inferences on (1) the relationship between the feedback they received with the task they had performed, (2) the joint representation of state and color belief for the current trial that led to their task belief, and finally (3) their current state belief that was used as an input for integration. This updating process thus requires the evaluation of multiple sources of information. Regions involved in updating must track current state and color beliefs (inputs) and evaluate the overall uncertainty of these inputs and the joint representation that influenced the current task belief, to determine the extent of updating required following a prediction error, which signals the need for updating.

To assess updating processes, we investigated brain regions that parametrically modulated their activity in response to both task and state prediction errors following feedback onset (Fig. **9A**). The assumption here is that incorrect feedback will first trigger a task prediction error, which will then trigger a re-evaluation of the integrated product and the inputs to integration (color and state belief). If subjects received incorrect feedback and they were highly certain in their color belief, then we would expect a state prediction error to arise from this backwards evaluation of information. To further investigate state updating processes, we also examined regions with activity profiles that fluctuated with the magnitude of trial-to-trial changes in state belief, which we refer to as the state derivative (Fig. **9D**). These state derivative analyses were time-locked to the cue onset on trial N in order to investigate updating processes occurring following feedback on the previous trial (trial N-1). Color updating processes were not evaluated because updating and encoding processes could not be dissociated due to the experimental design.

2.3.5.1 Prediction Errors: When assessing prediction errors, we focused on state and task prediction errors while controlling for general error processing (i.e., whether feedback was correct or incorrect). Using feedback-locked data, we predicted that voxels in the mPFC would show greater BOLD responses to larger state and task prediction errors, based on prior research linking the mPFC to conflict and error monitoring (Botvinick, Braver, Carter, & Cohen, 2001; Freund, Bugg, & Braver, 2021; Jiang, Wagner, & Egner, 2018; Rutledge, Dean, Caplin, & Glimcher, 2010; Schuck et al., 2016). As expected, both state and task prediction errors showed greater activation in mPFC (Fig. **9B-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. Although there was considerable overlap between state and task prediction errors, slightly stronger effects

were observed for task as compared to state prediction errors. This is expected given that (1) feedback was directly related to the task performed and (2) a task prediction error would trigger re-evaluation of state and color belief, which can in turn generate a state prediction error (depending on the degree of state uncertainty) that then leads to state belief updating.

2.3.5.2 State Derivative: We expected to find greater activity profiles in OFC following greater changes in state belief from trial N-1 to trial N given that the OFC is proposed to track state information (Moneta et al., 2024; Nassar et al., 2019; Schuck et al., 2016). Additionally, based on prior research suggesting subcortical structures, such as the basal ganglia, are involved in working memory updating, we expected these same structures to show a greater BOLD magnitude following a greater change in state belief. Presumably this would reflect greater updating costs related to updating the state information being tracked in working memory. In line with these expectations, we found that regions in OFC and the basal ganglia (caudate nucleus) showed a greater BOLD magnitude with greater state derivative (Fig. **9E**).

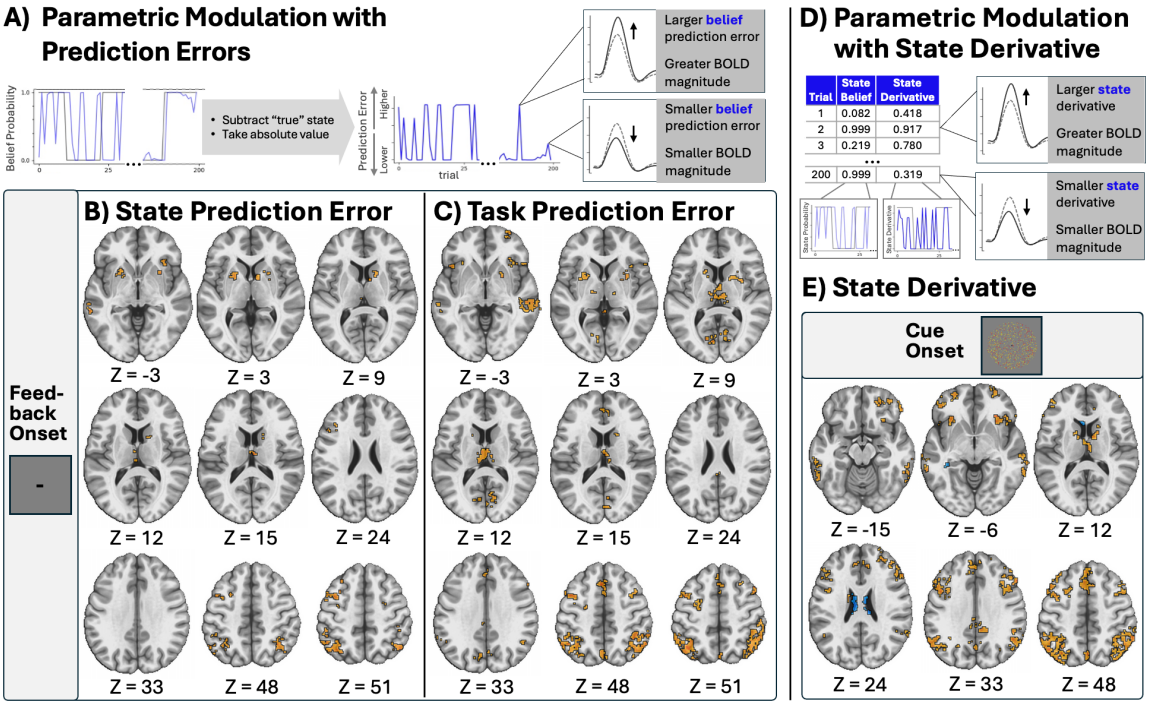

Figure 9. (**A**) Depiction of parametric modulation analyses with state and task prediction errors, and state derivative. Prediction error calculated by subtracting the "true" state/task from the model estimated state/take and taking the absolute values. Parametric modulation results with state (**B**) 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).

2.3.5.3 Functional Connectivity: Given that we also saw frontoparietal hub regions, such as the IPS, showing a greater BOLD magnitude with a greater prediction error and state derivative, we conducted an additional functional connectivity analysis to examine how entropy influences prediction error processes. Consistent with our prior

observation that entropy interacts with task output, we hypothesized that higher entropy would similarly modulate prediction error processing at feedback. That is based on the premise that prediction errors arise only when the system has knowledge of both the expected outcome and its associated uncertainty.

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 examine how entropy levels at task execution during the probe epoch influence prediction errors elicited during the feedback epoch, and (2) between beta estimates from ROIs during the feedback epochs. 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 (entropy) was associated with a significant positive modulation of functional connectivity between source entropy measured during probe epochs) and state prediction error during feedback, t=2.057, *p*=.040 (Figure **10A-B**). When connectivity was computed using only feedback epochs, no significant modulation was observed, t=-0.856, *p*=.392 (Figure **10C-D**). These results further support our interpretation that top-down influences utilize integrated uncertainty to bias task-related prediction error processing.

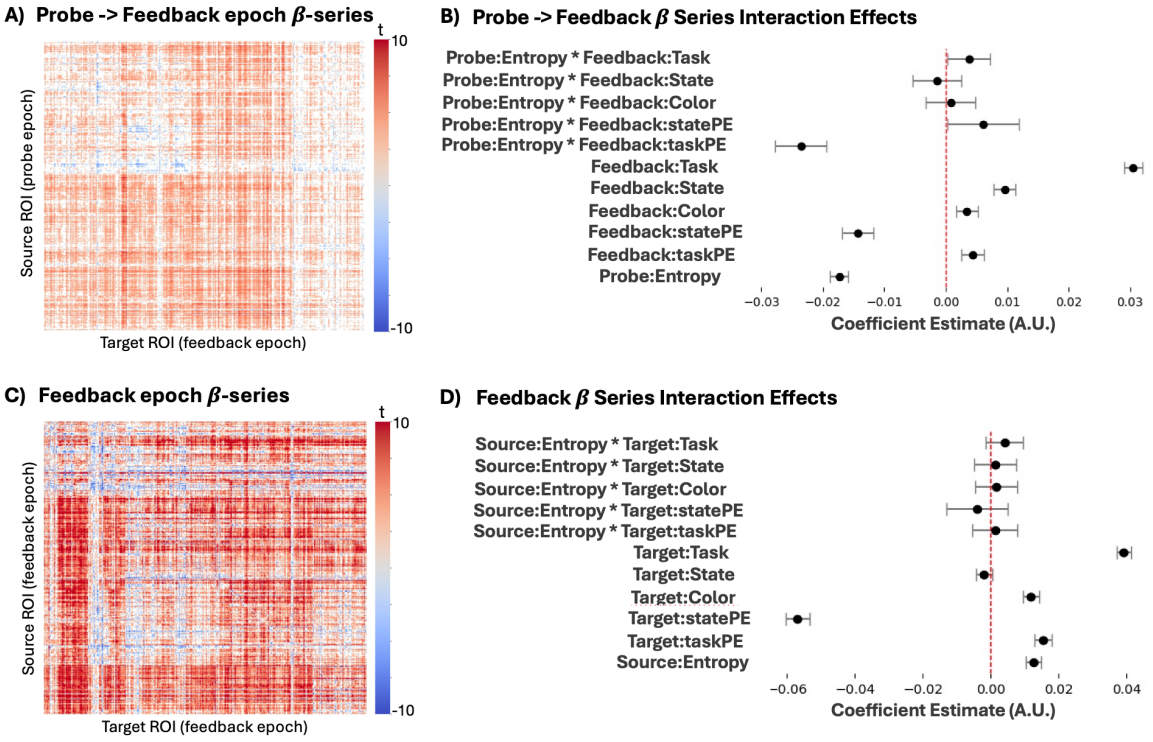

Figure 10. Beta-series connectivity matrices of entropy interaction for the cue-to-probe (**A**) epochs and feedback (**C**) epochs, displaying T-values from a mixed-effects regression analysis. Each cell represents the effect of a "source" ROI (rows) on a "target" ROI (columns). Warmer colors indicate higher positive T-values, while cooler colors indicate more negative T-values. **(B&D)** 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 supplement tables 4-5. (**A**) When using probe epochs as the source and feedback epochs as the target, higher

source entropy (at probe epochs) significantly modulates connectivity with target 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 error processing at feedback, when beliefs can be evaluated and updated. (**C**) When using only feedback epochs there were no significant interactions between entropy and prediction errors.

Discussion

Our primary goal was to determine the neural and computational mechanisms that support cognitive integration. Specifically, how internally maintained information (latent beliefs or expectations) is combined with externally presented sensory inputs to guide behavior. To address this question, we developed a Bayesian model that captures the full transformation from inputs to outputs by specifying: (1) the representational structure of task-relevant inputs, (2) the formation of an integrated representation (the joint probability distribution) that encodes the combined uncertainty of these inputs (quantified as entropy), (3) the probabilistic task belief from this joint representation, and (4) the updating mechanisms driven by prediction errors and trial-to-trial state representational changes. In our experiment, participants tracked a non-observable state and combined this state with color cues to choose between either the face or scene discrimination tasks, inferring error sources from non-informative feedback by evaluating their certainty in each input and in the integrated product. Our computational model provided trial-level estimates of both the categorical decision (the distribution's peak) and uncertainty (its width) for each representation. By integrating these estimates with fMRI—using decoding, parametric modulation, and connectivity analyses—we elucidated comprehensive, mechanistic account of cognitive integration.

Our first aim was to identify the neural substrates representing these distinct sources of information that needed to be integrated. Probabilistic decoding analyses revealed distinct regions encoded information about the two inputs for integration, color and state belief. Consistent with our predictions and in line with previous research, information from external, perceptual cue (color belief) was primarily represented in regions typically associated with functional networks such as the FPN, DAN, and VAN (Cole et al., 2013; Corbetta et al., 1998), while information from internally maintained/tracked information (state belief) was primarily represented in regions typically associated with the DMN (Konishi, McLaren et al. 2015, Vatansever, Menon et al. 2015, Nassar, McGuire et al. 2019). We did not find a cluster in OFC as we had expected with our more stringent a priori defined thresholds. However, we did find a small cluster with a less stringent threshold, which suggests the lack of a cluster in OFC could be due to a lack of power in the present study.

In short, these results indicated that external, perceptually grounded inputs and internally tracked, latent information were encoded in largely distinct neural substrates. This finding that the observable and non-observable inputs were decodable from distinct regions and networks further underscores the importance of investigating how our brains integrate these distinctly represented sources of information. Our computational model posits that integration is achieved by forming a joint representation that the combination of inputs, color and state. Therefore, the specific combination of color and state, which we refer to as the joint probability, was used to investigate which regions might encode the integrated representation.

Multinomial logistic regression decoding analyses revealed that the combination of inputs—represented by a 2×2 probability matrix—accurately predicted the state-color combination with the highest probability that predicts task output. Some of these regions appear to overlap with regions where state or color (input) belief was decoded. Moreover, the output (task) belief was decodable from a subset of regions where the state-color joint probability could be decoded from. Notably, FPN hub regions, such as MFG and IPS, as well as sensory motor network regions and task-specific regions such as the FFA/PPA. This finding supports the idea that the integration of inputs, resulting in the strongest joint probability, can predict the task belief.

A critical component of our framework is that the joint representation encapsulates the combined uncertainty of each input – specifically, the entropy of the joint distribution reflected the overall uncertainty related to integration. We hypothesized that this combined uncertainty acts as a biasing signal to alter connectivity patterns to guide sensory inputs into a single, integrated task output. We tested this in two ways. First, we examined uncertainty at multiple levels—individual input uncertainty (color and state), integrated uncertainty (entropy), and output uncertainty (task belief)—by using these estimates as parametric modulators of the BOLD response and tested if they involved similar neural substrates. Second, we performed novel connectivity analyses to assess how integrated uncertainty influences the output (task) belief.

Univariate parametric modulation analyses revealed a striking overlap in regions that modulated their activity with all uncertainty measures. Of note, regions such as the MFG and IPS have been found to be hub regions connected to many large-scale networks (Bertolero et al., 2018; Cocuzza, Ito, Schultz, Bassett, & Cole, 2020; Cole et al., 2013; Gratton et al., 2018). A finding we replicated with network hub analyses in the current study. Critically, regions that uniquely modulated with either state or color uncertainty also showed corresponding fluctuations in entropy and task uncertainty. This convergence suggests that, instead of resolving uncertainty at the input level and considering only the categorical decision, the human brain preserves the full probabilistic information of both inputs and processes them within a common set of frontoparietal regions. Additionally, it further suggests that the integrated representation is not limited to a single brain region; rather it is represented across a wide range of networks, specifically in hub regions. Therefore, integration might happen through a biasing mechanism that resolves uncertainty to select and output the "optimal" option. If that is the case, the FPN is the most likely candidate given (1) previous research showing that FPN hub regions change their global connectivity to implement control (Cole, 2017; Cole, 2013) and (2) the overlap between our own network hub analysis and regions fluctuating with entropy/uncertainty. In other words, FPN might be functioning like a low dimensional readout system that detects and combines uncertainty estimates from relevant information sources. Then, the FPN resolves the competition that arises due to this overall uncertainty and acts as a biasing signal to downstream regions to drive the output of a single task belief from this joint representation.

To directly assess the abovementioned prediction, we tested the influence of integrated uncertainty (entropy) on the output (task) belief via a set of novel connectivity analyses. 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 the entropy measure included as a moderator. We found that, during integration, increased uncertainty enhanced connectivity related to processing external sensory information (i.e., color beliefs) while decreasing connectivity with regions encoding task output. In contrast, following integration, the connectivity pattern shifted such that

integrated uncertainty primarily influenced connectivity with regions encoding the integrated task output, indicative of a top-down biasing interaction. Additionally, when looking at cue-to-probe connectivity patterns, which allowed us to infer a temporally directional effect, we found that integrated uncertainty directly influenced connectivity with regions encoding the integrated task output. In other words, these connectivity analyses further suggest that entropy, a measure of the overall uncertainty in integration, influences the output of integration through a top-down biasing mechanism.

The results discussed so far have suggested that entropy acts as a biasing signal to alter connectivity patterns to output a single task belief. However, our account of integration would be incomplete without considering the updating processes that rely on this information. Given state belief had no explicit external cue, this belief had to be tracked and updated on a trial-by-trial basis. Accordingly, we investigated the updating mechanisms that adjusted state belief using feedback-driven prediction errors and information from the other input (color belief), the integrated representation, and the task output (task belief). We predicted that regions tracking integration would show activity fluctuations proportional to these prediction errors. Moreover, regions involved specifically in updating the internally maintained state belief should fluctuate their activity patterns based on the magnitude of the change in state belief across trials because a greater change in state belief should impose greater updating demands.

Our 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. When updating the internally maintained and tracked state belief, updating processes need information about the other input (color belief), integrated product, and output (task belief), all of which will need to be maintained in working memory for the duration of the trial. Therefore, updating processes in our study appeared to operate in (1) regions that fluctuated with input, joint representation, and output uncertainty, and (2) distinct cortical and subcortical regions that contained no information about input, integrated representation, or output beliefs/uncertainties. This suggests an updating system that relies on a biasing signal similar to the integration system. In other words, the frontoparietal hub regions utilize the integrated uncertainty to bias regions involved in the detection of violations in expectations (prediction errors) in order to drive the 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 (Ma, Beck, Latham, & Pouget, 2006; van Bergen, Ma, Pratte, & Jehee, 2015). However, decision making requires a winner-take-all process where only the most probable option is selected (Vul, Hanus, & Kanwisher, 2009). Therefore, to decide on a task or action, the brain must maintain and integrate probabilistic representations that ultimately select a single outcome to guide behavior. Our computational framework suggests that probabilistically maintained representations are integrated into a joint distribution that produces a conditional probabilistic distribution from which the most likely correct action plan is selected. Although this output remains probabilistic—facilitating subsequent updating—it drives motor plans in a winner-take-all manner. Therefore, the inputs, integrated product, and output all need to be maintained probabilistically across a diverse set of regions, enabling targeted updates in response to the unique updating demands of each representation.

Furthermore, these results suggest that the overlapping regions modulating their activity with uncertainty and entropy 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 the input information. Recent work by Li, Sprague, Yoo, Ma, and Curtis (2024) found that the same regions that track overall working memory resource allocation may leverage uncertainty signals to drive targeted updating and resolve ambiguity of working memory representations in a winner-take-all manner. We found similar results with prediction errors and updating processes; that is to say, greater entropy—greater integrated uncertainty—modifies connectivity patterns to enhance connectivity related to state prediction errors to drive targeted updating of internally maintained state information. In other words, uncertainty appears to be a critical mechanism by which the brain continuously integrates probabilistic information, ensuring adaptive control over behavior.

In order to investigate the integration of external and internal sources of information, there was the inevitable limitation of the certainty in the observable color information impacting the uncertainty in the non-observable state information. Participants could only be certain in a state change when they were certain in their color belief on a given trial where they received incorrect feedback. This limitation could be overcome by using two observable sources of information; however, this would fail to address the critical question outlined in this study, which was how the brain integrates observable and non-observable information. This limitation could also be overcome with the introduction of more inputs, with only a subset of inputs being relevant on any given trial. However, this will inevitably increase the complexity of the paradigm and muddy interpretations of results.

In summary, our work introduces a computational framework that describes how internally maintained and externally cued information is integrated to guide 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. Our results revealed that internal and external inputs were represented in distinct brain structures, which then converged in frontoparietal hub regions that dynamically modulated connectivity based an integrated uncertainty estimate. The task output was primarily encoded in premotor and motor areas, reflecting the final selection process. Finally, updating mechanisms recruited both frontoparietal and subcortical structures (basal ganglia, thalamus) to update internal representations based on feedback. Our work answers a fundamental question of how the human neurocognitive system integrates and updates sources of 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 this Bayesian model, for trial i, the probabilistic state belief (S_i) and probabilistic color belief (C_i) are integrated into a 2x2 joint distribution, which was used to infer the probabilistic task belief (T_i).

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

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.

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

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 (Behrens, Woolrich, Walton, & Rushworth, 2007; Jiang, Heller, & Egner, 2014).

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).

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

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):

(4) Equation $$H = -1 * \Sigma\, p_i * \log(p_i)$$

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 (Esteban et al., 2019) to reduce noise and transform data from subject native space to the ICBM 152 Nonlinear Asymmetrical template version 2009c for group analysis (Fonov et al., 2011). 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 matter (WM) noise components from the component-based noise correction procedure (Behzadi, Restom, Liau, & Liu, 2007). These nuisance regressors were entered in the regression model to reduce the influences from noise and artifacts. We did not perform any spatial smoothing when preprocessing the data.

3.8 Statistical Analysis

3.8.1 Mixed Effects Models Predicting Response Times: To examine how trial-by-trial variations in cognitive integration influence behavior, we conducted two separate mixed effects models predicting participant response times (RTs). The models used either (1) the

entropy of the joint distribution or (2) the uncertainty of the task belief, which was the output of the joint distribution. Both models included random intercepts for subjects. Response times faster than 200 milliseconds and slower than 3500 milliseconds (offset of probe image) were excluded from analyses. The first trial of each block was also excluded. The entropy of the joint distribution was winsorized to address outliers (values in the bottom 5% and top 95% were replaced with the values at 5th or 95th percentile, respectively). Winsorized entropy and task uncertainty values were approximately normally distributed and homoscedastic. We performed two separate mixed effects regressions due to the high correlation between entropy and task uncertainty ($r = .76$). We predicted that both greater entropy and greater task uncertainty would predict slower RTs.

3.8.2 Parametric Modulation: To relate our model parameters to fMRI data, we conducted parametric modulation to identify brain regions where BOLD response amplitudes vary systematically with key parameters from our computational model: (1) the uncertainty of the model predicted state belief, color belief, and task belief, (2) prediction errors and derivatives of task and state beliefs, and (3) the entropy of the joint distribution. These analyses aimed to capture the representational structure, its uncertainty estimates, and computational processes underlying cognitive integration. This 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 (Cox, 1996). 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 (Cox, 1996). Decoding was performed via linear regression (Python package scikit-learn; (Pedregosa, 2011)), 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 (Cox, 1996). We then conducted a multinomial logistic regression (Python package scikit-learn; (Pedregosa, 2011)), 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-color1, state1-color2, state2-color1, state2-color2), chance level decoding performance was set at 0.25. Finally, we conducted mass univariate t-tests that tested if these decoding accuracy values across subjects were significantly different from 0.25. To correct for multiple comparisons, we used 3dClustSim with the residuals from a voxel-wise 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).

3.8.5 Connectivity Analysis: We conducted a customized connectivity analysis to examine task-related connectivity patterns between ROIs during the cue, probe, and feedback epochs, and to assess how trial-by-trial entropy values modulate these patterns. Our approach builds on the beta-series connectivity framework (Rissman, Gazzaley, & D'Esposito, 2004) by first extracting single-trial beta estimates from LSS deconvolution using the Schaefer 400-ROI atlas (Schaefer et al., 2018), and adapted the generalized psychophysiology interaction framework by fitting a regression model to the trial-wise beta values (McLaren, Ries, Xu, & Johnson, 2012). Note that this same set of LSS single-trial beta estimates were used for the decoding analyses.

For each subject, we computed pairwise regression models for all ROI pairs (excluding self-coupling). In these models, the trial-wise beta values from a "target" ROI was predicted by the beta values from a "source" ROI, and its interaction with the trial-by-trial entropy estimates. To control for outliers, independent variables were winsorized and centered before interaction terms were computed. We then fitted ordinary least squares regression models within subjects, and combined these results in a mixed-effects framework across subjects, using random intercepts to account for subject variability. This regression approach yielded estimates of both the main connectivity effect and its interaction with integrated uncertainty. The interaction effect indicates how the functional connectivity of one ROI on another ROI varies with the entropy of the joint 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 (Gratton, Nomura, Perez, & D'Esposito, 2012; Hwang, Shine, Bruss, Tranel, & Boes, 2021). This result is implemented using data and methods developed in Reber et al. (2021), and briefly described here.

The PC value for each ROI, PC_i is defined as

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

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 voxelwise value of 0.65 in order to visualize regions with connector hub-like properties.

References

[1] Rangel, B. O., Hazeltine, E., & Wessel, J. R. (2023). Lingering Neural Representations of Past Task Features Adversely Affect Future Behavior. *J Neurosci, 43*(2), 282-292. doi:10.1523/JNEUROSCI.0464-22.2022

[2] Cellier, D., Petersen, I. T., & Hwang, K. (2022). Dynamics of Hierarchical Task Representations. *J Neurosci, 42*(38), 7276-7284. doi:10.1523/JNEUROSCI.0233-22.2022

[3] Gratton, C., Sun, H., & Petersen, S. E. (2018). Control networks and hubs. *Psychophysiology, 55*(3). doi:10.1111/psyp.13032

[4] Power, J. D., Schlaggar, B. L., Lessov-Schlaggar, C. N., & Petersen, S. E. (2013). Evidence for hubs in human functional brain networks. *Neuron, 79*(4), 798-813. doi:10.1016/j.neuron.2013.07.035

[5] van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends Cogn Sci, 17*(12), 683-696. doi:10.1016/j.tics.2013.09.012

[6] Sporns, O., & Betzel, R. F. (2016). Modular Brain Networks. *Annu Rev Psychol, 67*, 613-640. doi:10.1146/annurev-psych-122414-033634

[7] Bertolero, M. A., Yeo, B. T., & D'Esposito, M. (2015). The modular and integrative functional architecture of the human brain. *PNAS*. doi:10.1073/pnas.1510619112Downloaded

[8] Bertolero, M. A., Yeo, B. T. T., Bassett, D. S., & D'Esposito, M. (2018). A mechanistic model of connector hubs, modularity and cognition. *Nat Hum Behav, 2*(10), 765-777. doi:10.1038/s41562-018-0420-6

[9] Cole, M. W., Reynolds, J. R., Power, J. D., Repovs, G., Anticevic, A., & Braver, T. S. (2013). Multi-task connectivity reveals flexible hubs for adaptive task control. *Nat Neurosci, 16*(9), 1348-1355. doi:10.1038/nn.3470

[10] Vatansever, D., Menon, D. K., Manktelow, A. E., Sahakian, B. J., & Stamatakis, E. A. (2015). Default Mode Dynamics for Global Functional Integration. *J Neurosci, 35*(46), 15254-15262. doi:10.1523/JNEUROSCI.2135-15.2015

[11] Badre, D., & D'Esposito, M. (2007). Functional Magnetic Resonance Imaging Evidence for a Hierarchical Organization of the Prefrontal Cortex. *Journal of Cognitive Neuroscience, 19*(12), 2082-2099.

[12] Badre, D., Hoffman, J., Cooney, J. W., & D'Esposito, M. (2009). Hierarchical cognitive control deficits following damage to the human frontal lobe. *Nat Neurosci, 12*(4), 515-522. doi:10.1038/nn.2277

[13] Nee, D. E., & D'Esposito, M. (2016). The hierarchical organization of the lateral prefrontal cortex. *Elife, 5*, e12112. doi:10.7554/eLife.12112

[14] Collins, A. G., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG uncovers latent generalizable rule structure during learning. *J Neurosci, 34*(13), 4677-4685. doi:10.1523/JNEUROSCI.3900-13.2014

[15] Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron, 83*(1), 238-251. doi:10.1016/j.neuron.2014.05.014

[16] Hwang, K., Shine, J. M., & D'Esposito, M. (2019). Frontoparietal Activity Interacts With Task-Evoked Changes in Functional Connectivity. *Cereb Cortex, 29*(2), 802-813. doi:10.1093/cercor/bhy011

[17] Moneta, N., Grossman, S., & Schuck, N. W. (2024). Representational spaces in orbitofrontal and ventromedial prefrontal cortex: task states, values, and beyond. *Trends Neurosci, 47*(12), 1055-1069. doi:10.1016/j.tins.2024.10.005

[18] Nassar, M. R., McGuire, J. T., Ritz, H., & Kable, J. W. (2019). Dissociable Forms of Uncertainty-Driven Representational Change Across the Human Brain. *J Neurosci, 39*(9), 1688-1698. doi:10.1523/JNEUROSCI.1713-18.2018

[19] Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a Cognitive Map of State Space. *Neuron, 91*(6), 1402-1412. doi:10.1016/j.neuron.2016.08.019

[20] Corbetta, M., Akbudak, E., Conturo, T. E., Synder, A. Z., Ollinger, J. M., Drury, H. A., . . . Shulman, G. L. (1998). A common network of functional areas for attention and eye movements. *Neuron, 21*, 761-773.

[21] Cole, M. W., Ito, T., Cocuzza, C., & Sanchez-Romero, R. (2021). The Functional Relevance of Task-State Functional Connectivity. *J Neurosci, 41*(12), 2684-2702. doi:10.1523/JNEUROSCI.1713-20.2021

[22] Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. *Neuroimage, 23*(2), 752-763. doi:10.1016/j.neuroimage.2004.06.035

[23] Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X. N., Holmes, A. J., . . . Yeo, B. T. T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cereb Cortex, 28*(9), 3095-3114. doi:10.1093/cercor/bhx179

[24] Reber, J., Hwang, K., Bowren, M., Bruss, J., Mukherjee, P., Tranel, D., & Boes, A. D. (2021). Cognitive impairment after focal brain lesions is better predicted by damage to structural than functional network hubs. *Proc Natl Acad Sci U S A, 118*(19). doi:10.1073/pnas.2018784118

[25] Botvinick, M., Braver, T. S., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review, 108*. doi:10.1037//0033-295X.I08.3.624

[26] Freund, M. C., Bugg, J. M., & Braver, T. S. (2021). A representational similarity analysis of cognitive control during color-word stroop. *bioRxiv*. doi:10.1101/2020.11.22.392704

[27] Jiang, J., Wagner, A. D., & Egner, T. (2018). Integrated externally and internally generated task predictions jointly guide cognitive control in prefrontal cortex. *Elife, 7*. doi:10.7554/eLife.39497

[28] Rutledge, R. B., Dean, M., Caplin, A., & Glimcher, P. W. (2010). Testing the reward prediction error hypothesis with an axiomatic model. *J Neurosci, 30*(40), 13525-13536. doi:10.1523/JNEUROSCI.1747-10.2010

[29] Cocuzza, C. V., Ito, T., Schultz, D., Bassett, D. S., & Cole, M. W. (2020). Flexible Coordinator and Switcher Hubs for Adaptive Task Control. *J Neurosci, 40*(36), 6949-6968. doi:10.1523/JNEUROSCI.2559-19.2020

[30] Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat Neurosci, 9*(11), 1432-1438. doi:10.1038/nn1790

[31] van Bergen, R. S., Ma, W. J., Pratte, M. S., & Jehee, J. F. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nat Neurosci, 18*(12), 1728-1730. doi:10.1038/nn.4150

[32] Vul, E., Hanus, D., & Kanwisher, N. (2009). Attention as inference: selection is probabilistic; responses are all-or-none samples. *J Exp Psychol Gen, 138*(4), 546-560. doi:10.1037/a0017352

[33] Li, H., Sprague, T. C., Yoo, A. H., Ma, W. J., & Curtis, C. E. (2024). Neural mechanisms of resource allocation in working memory.

[34] Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci, 10*(9), 1214-1221. doi:10.1038/nn1954

[35] Jiang, J., Heller, K., & Egner, T. (2014). Bayesian modeling of flexible cognitive control. *Neurosci Biobehav Rev, 46 Pt 1*, 30-43. doi:10.1016/j.neubiorev.2014.06.001

[36] Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., . . . Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat Methods, 16*(1), 111-116. doi:10.1038/s41592-018-0235-4

[37] Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., & Collins, D. L. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage, 54*. doi:10.1016/j.neuroimage.2010.07.033314

[38] Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage, 37*(1), 90-101. doi:10.1016/j.neuroimage.2007.04.042

[39] Cox, R. W. (1996). AFNI Software for analysis and visualization of functional magnetic resonance neuroimages.

[40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. . (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research, 12*, 2825-2830.

[41] McLaren, D. G., Ries, M. L., Xu, G., & Johnson, S. C. (2012). A generalized form of context-dependent psychophysiological interactions (gPPI): a comparison to standard approaches. *Neuroimage, 61*(4), 1277-1286. doi:10.1016/j.neuroimage.2012.03.068

[42] Gratton, C., Nomura, E. M., Perez, F., & D'Esposito, M. (2012). Focal brain lesions to critical locations cause widespread disruption of the modular organization of the brain. *Journal of Cognitive Neuroscience, 24*.

[43] Hwang, K., Shine, J. M., Bruss, J., Tranel, D., & Boes, A. (2021). Neuropsychological evidence of multi-domain network hubs in the human thalamus. *Elife, 10*. doi:10.7554/eLife.69480

Acknowledgments

Research reported here was supported by a National Institute of Mental Health grant R01MH122613 (KH)

Research reported here was supported by Iowa Neuroscience Institute 36

This work was conducted on an MRI instrument funded by 1S10OD025025-01.

Author contributions are as follows.

- Conceptualization: SCL, KH, JJ
- Methodology: SCL, KH, JJ
- Investigation: SCL, HM, 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.

Figures and Tables

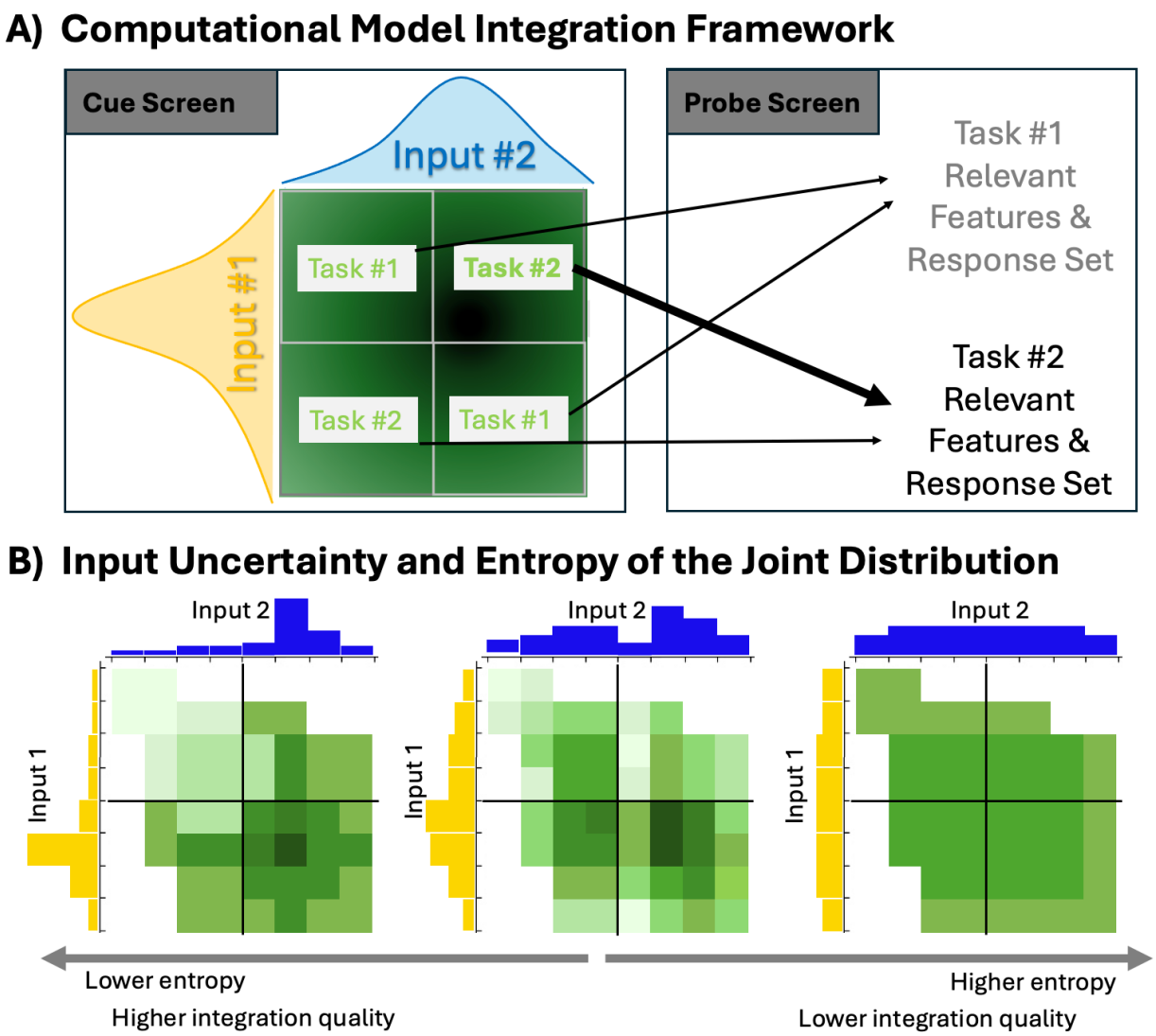

Fig. 1. Entropy and Integration. (**A**) Depiction of integration of information into a joint distribution and its corresponding conditional probability distribution that predicts the specific task or decision to perform. (**B**) Depiction of entropy estimate of the joint distribution of two input variables: state and color. Low entropy resulting from low uncertainty in both inputs. Intermediate entropy due to moderate uncertainty in one or both inputs, whereas high entropy is due to very high uncertainty in the inputs.

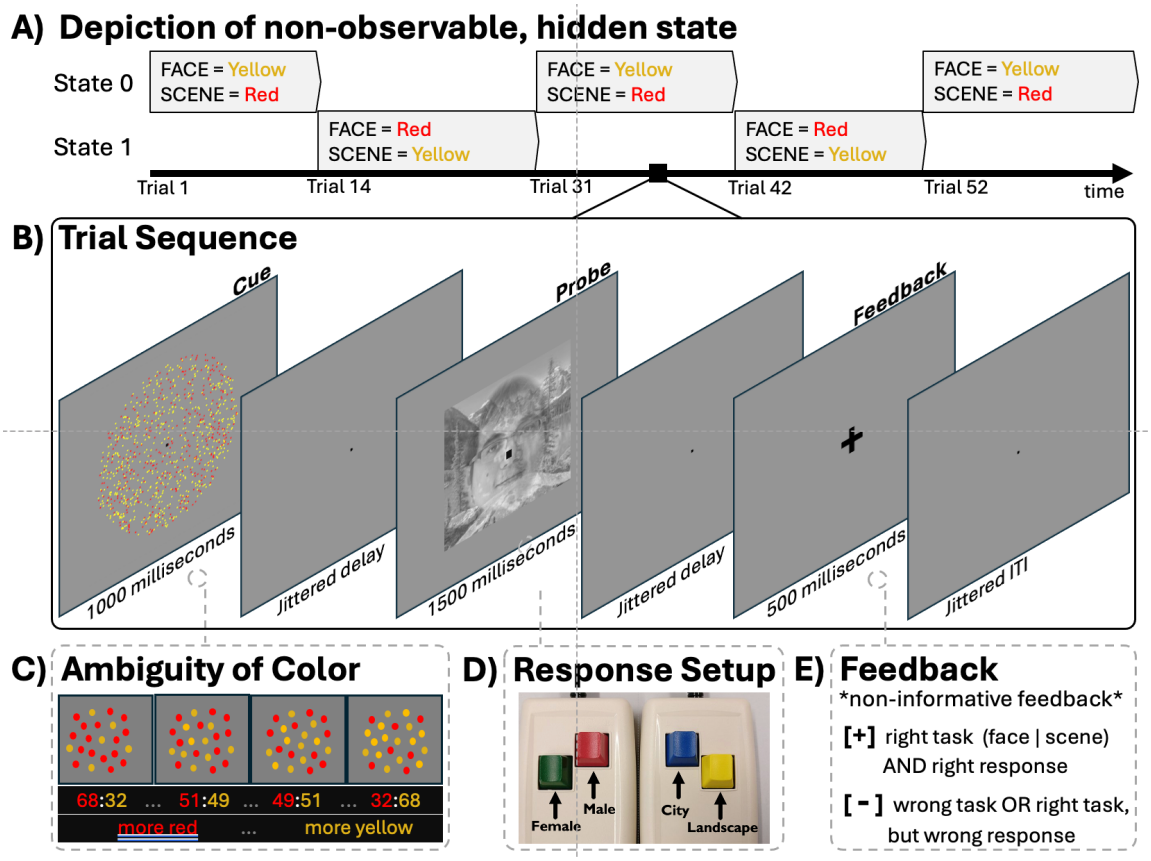

Fig. 2. Experimental Design. (**A**) Depiction of non-observable state that changes every 10-20 trials. **(B)** Trial sequence example: cue screen (1000ms), jittered delay (range: 0.5-10.5 seconds, average=3.75 seconds), probe screen (1500ms), jittered delay (range: 0.5-10.5 seconds, average=3.75 seconds), 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.

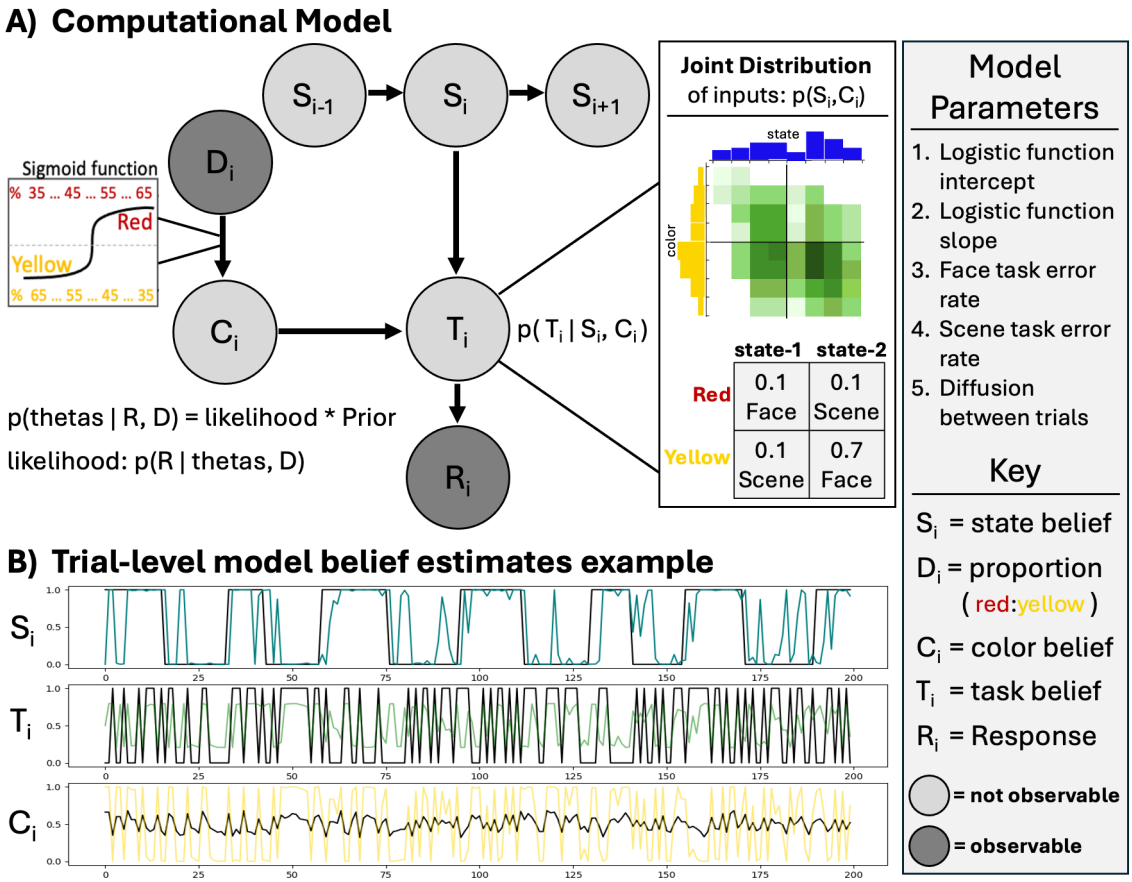

Fig. 3. Computational Model. (**A**) Computational (Bayesian generative) model diagram. Observable information (R_i and D_i) used to infer three latent variables (S_i, C_i, and T_i). (**B**) Trial-level model estimated "beliefs" from one example subject. Black lines: true state, task, or color. Blue,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

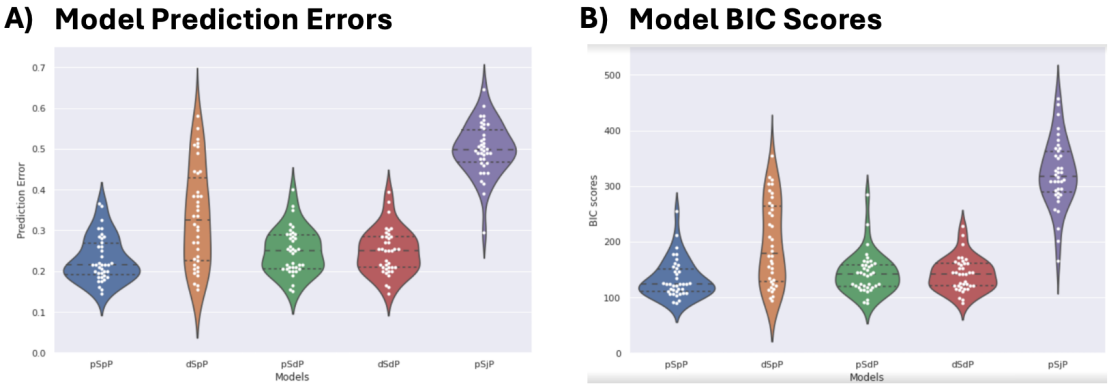

Fig. 4. Integration Uncertainty and RTs. (**A**) Prediction error (percent of trials) of each computational model. Each dot represents the prediction accuracy for an individual subject. (**B**) BIC scores for each computational model. Each dot represents the BIC score for an individual subject.

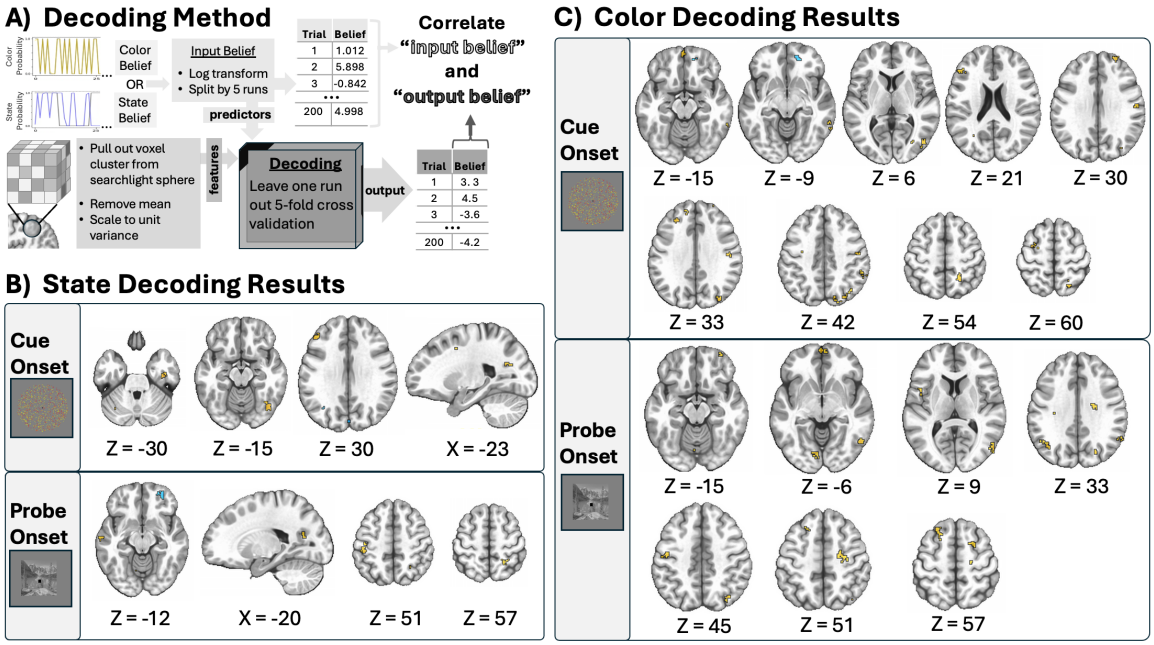

Fig. 5. Input Decoding. (**A**) Depiction of decoding method for color and state belief. Decoding results for color belief (**C**) and state (**B**) belief with cue and probe epochs. Results based on minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold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

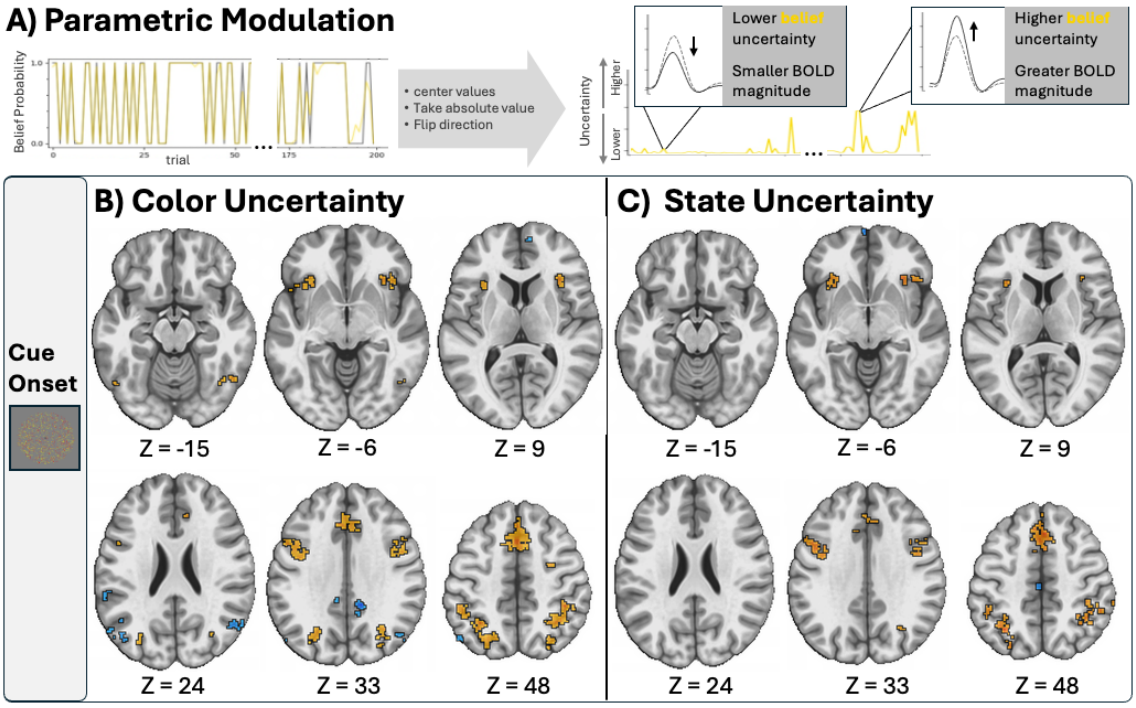

Fig. 6. Input Parametric Modulation. (**A**) Depiction of parametric modulation analyses for color uncertainty and state uncertainty. Uncertainty calculated by centering color/state belief, taking the absolute value, and flipping the direction so that a larger value reflects greater uncertainty. Color (**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).

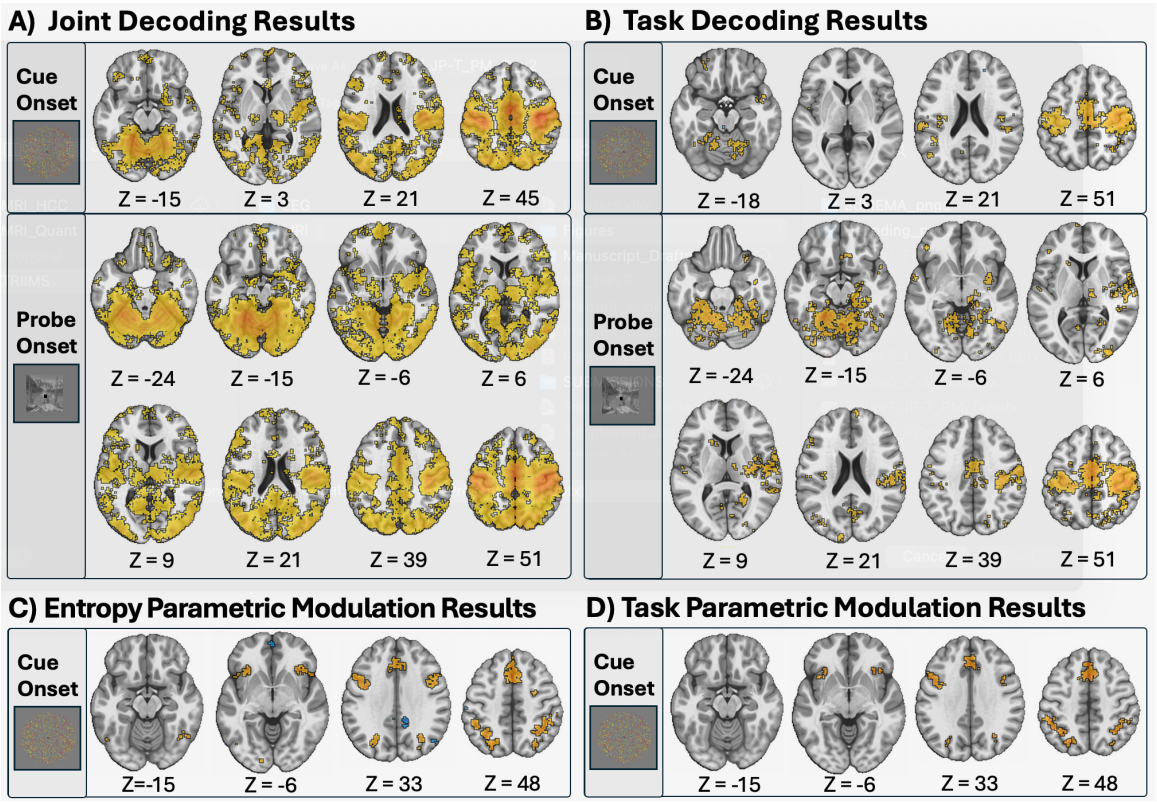

Fig. 7. Integration and Output Decoding and Parametric Modulation. Decoding results (**A-B**) for joint probability (**A**) and task belief (**B**) with cue and probe epochs based on minimum cluster size of 21 voxels (cluster p-value threshold of 0.01 and a voxel-level p-value threshold of .005). Parametric 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.

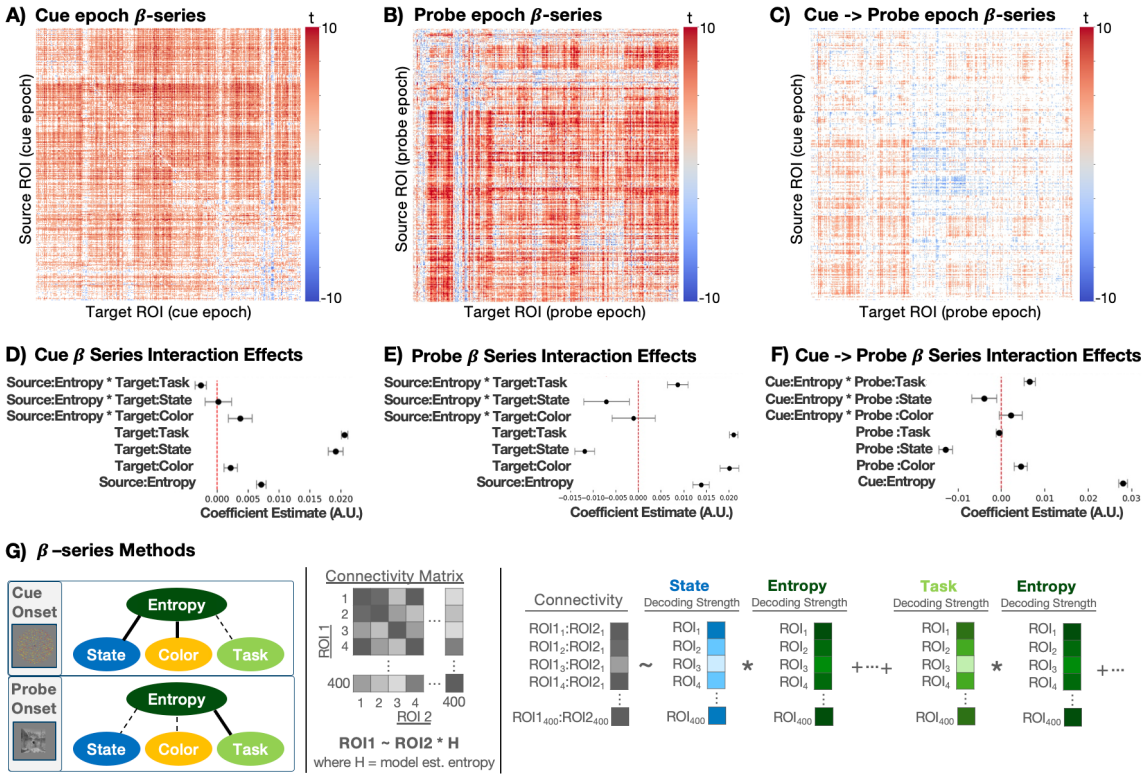

Fig. 8. Functional Connectivity and Integration. Beta-series connectivity matrices of entropy interaction for the cue (**A**) and probe (**B**) epochs as well as cue-to-probe (**C**) epochs, displaying T-values from a mixed-effects regression analysis. Each cell represents the effect of a "source" ROI (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. The face stimuli in the probe example is the corresponding author.

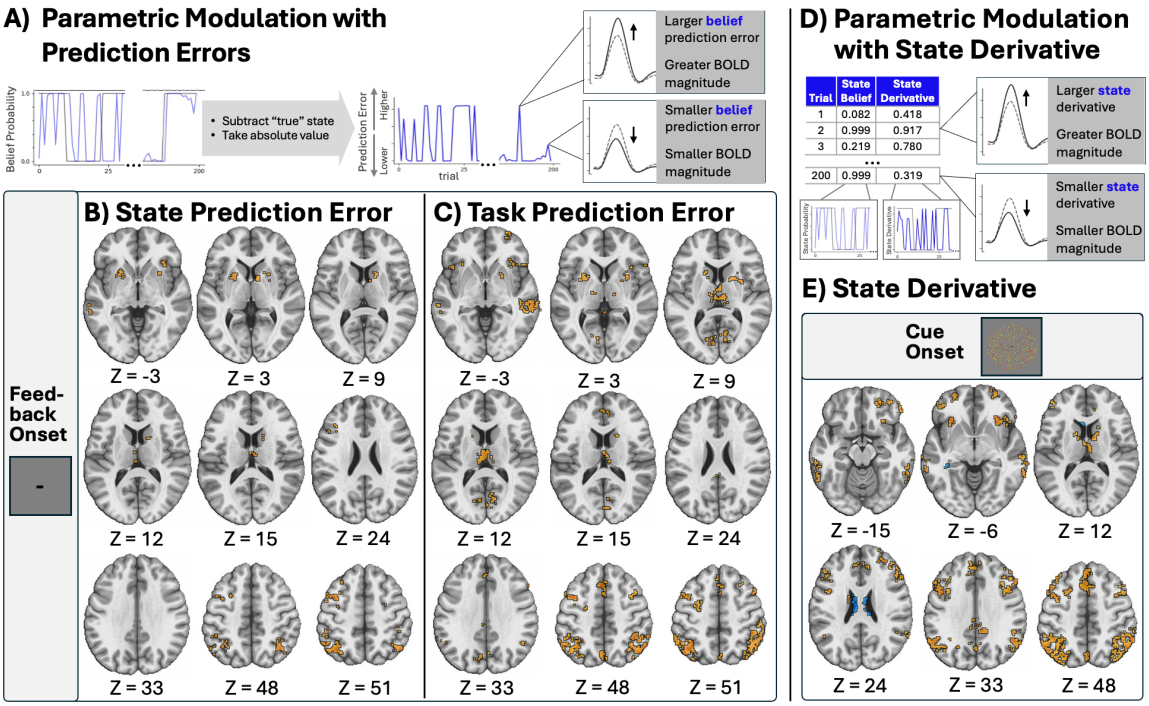

Fig. 9. Prediction Errors and Updating Parametric Modulation. (**A**) Depiction of parametric modulation analyses with state and task prediction errors, and state derivative. Prediction error calculated by subtracting the "true" state/task from the model estimated state/take and taking the absolute values. 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).

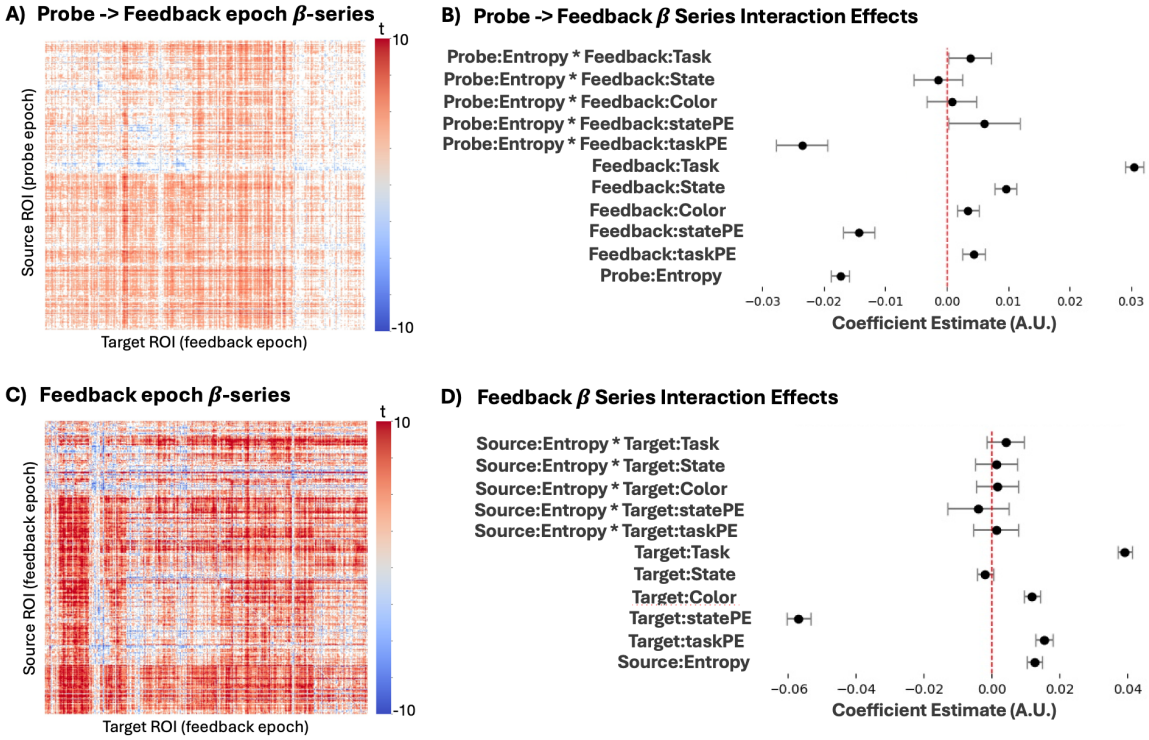

Fig. 10. Functional Connectivity and Prediction Errors. Beta-series connectivity matrices of entropy interaction for the cue-to-probe (**A**) epochs and feedback (**C**) epochs, displaying T-values from a mixed-effects regression analysis. Each cell represents the effect of a "source" ROI (rows) on a "target" ROI (columns). Warmer colors indicate higher positive T-values, while cooler colors indicate more negative T-values. **(B&D)** 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 supplement tables 4-5. (**A**) When using probe epochs as the source and feedback epochs as the target, higher source entropy (at probe epochs) significantly modulates connectivity with target 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 error processing at feedback, when beliefs can be evaluated and updated. (**C**) When using only feedback epochs there were no significant interactions between entropy and prediction errors.

Supplementary Materials

See associated supplement file