Behavioral/Cognitive

Neural Representations of Sensory Uncertainty and Confidence Are Associated with Perceptual Curiosity

Michael Cohanpour,[1,2] Mariam Aly,[3*] and Jacqueline Gottlieb[1,2,4*]

[1]Department of Neuroscience, Columbia University, New York, New York 10025, [2]Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, New York 10025, [3]Department of Psychology, Columbia University, New York, New York 10025, and [4]Kavli Institute for Brain Science, Columbia University, New York, New York 10025

Humans are immensely curious and motivated to reduce uncertainty, but little is known about the neural mechanisms that generate curiosity. Curiosity is inversely associated with confidence, suggesting that it is triggered by states of low confidence (subjective uncertainty), but the neural mechanisms of this link, have been little investigated. Inspired by studies of sensory uncertainty, we hypothesized that visual areas provide multivariate representations of uncertainty, which are read out by higher-order structures to generate signals of confidence and, ultimately, curiosity. We scanned participants (17 female, 15 male) using fMRI while they performed a new task in which they rated their confidence in identifying distorted images of animals and objects and their curiosity to see the clear image. We measured the activity evoked by each image in the occipitotemporal cortex (OTC) and devised a new metric of "OTC Certainty" indicating the strength of evidence this activity conveys about the animal versus object categories. We show that, perceptual curiosity peaked at low confidence and OTC Certainty negatively correlated with curiosity, establishing a link between curiosity and a multivariate representation of sensory uncertainty. Moreover, univariate (average) activity in two frontal areas—vmPFC and ACC—correlated positively with confidence and negatively with curiosity, and the vmPFC mediated the relationship between OTC Certainty and curiosity. The results reveal novel mechanisms through which uncertainty about an event generates curiosity about that event.

Key words: **confidence; curiosity; information seeking; noninstrumental; uncertainty; visual cortex**

Significance Statement

Curiosity motivates us to learn and explore. Traditional perspectives hypothesize that curiosity arises from variability in confidence, but the neural mechanisms by which this occurs have been difficult to evaluate. Here, we harness the human visual system to uncover a neural mechanism of curiosity. We show that a multivariate representation of certainty in the occipitotemporal cortex is transformed into a univariate representation of confidence in the prefrontal cortex to facilitate curiosity. These results illuminate how perceptual input is transformed by successive neural representations to ultimately evoke a feeling of curiosity—elucidating how and why we become curious to learn about diverse knowledge domains.

Received May 26, 2023; revised April 7, 2024; accepted June 18, 2024.

Author contributions: M.C. and J.G. conceptualized the experiment. M.C., M.A., and J.G. designed research; M.C. performed research; M.C., M.A., and J.G. contributed unpublished reagents/analytic tools; M.C. analyzed data; M.C., M.A., and J.G. wrote the paper.

The research described in this paper was supported by the National Institute of Mental Health as part of the National Research Service Award (Grant #1F31MH125589), and the Zuckerman Institute MR Seed Grant Award (Grant # CU-ZI-MR-S-0017) both awarded to M.C. We thank the Alyssano Group, Gottlieb Lab, Kriegeskorte Lab, Christopher Baldassano, Janet Metcalfe, and Yasmine El-Shamayleh for their valuable insight on this project; Ray Lee and Noreen Violante for their technical support with the MRI scanner; and Serra Favila, Heiko Schütt, and Javier Domínguez Zamora for their crucial revisions to the manuscript.

*M.A. and J.G. contributed equally to this work.

The authors declare no competing financial interests.

M. A.'s present address: Department of Psychology, University of California Berkeley, Berkeley, California 94720.

Correspondence should be addressed to Michael Cohanpour at mcohanpour@gmail.com or Mariam Aly at ma3631@columbia.edu or Jacqueline Gottlieb at jg2141@columbia.edu.

https://doi.org/10.1523/JNEUROSCI.0974-23.2024

Introduction

Humans are immensely curious—we are motivated to seek information even if this comes with no monetary rewards (Gottlieb and Oudeyer, 2018; Kidd and Hayden, 2015; van Lieshout et al., 2020). Curiosity compels us to explore, inquire, learn, and discover. What neural mechanisms underlie curiosity?

Recent studies examined curiosity using tasks in which participants are shown trivia questions and asked to rate their curiosity about receiving the answer. These studies have shown that curiosity ratings are encoded in brain regions implicated in motivation and reward (Kang et al., 2009; Gruber et al., 2014). Moreover, curiosity spurs cognitive actions, including anticipatory shifts of attention for gathering information (Baranes et al., 2015), and enhancement of memory for the

answer through functional interactions between the hippocampus and reward circuitry (Gruber et al., 2014; Murphy et al., 2021).

Despite these important advances, challenging open questions remain about the mechanisms that generate curiosity. How does the neural representation of an event—evoke a state of high or low curiosity about that event?

Converging evidence suggests that this link, between an internal representation and feelings of curiosity, is mediated by confidence or uncertainty. Studies of curiosity in which participants are also asked to rate their confidence about knowing the answer to a trivia question report that ratings of confidence and curiosity are inversely related—i.e., lower confidence is associated with higher curiosity (Kang et al., 2009; Baranes et al., 2015). This suggests that the representation of a trivia question may generate low or high uncertainty which, in turn, evokes high or low curiosity (Berlyne, 1954; Loewenstein, 1994; Golman and Loewenstein, 2018). However, the neural mechanisms linking uncertainty, confidence, and curiosity remain largely unknown, in large part because we have scant understanding of how the brain represents the uncertainty of semantic information that is indexed by trivia questions.

Here we sought to examine this question by capitalizing on the richer literature on sensory uncertainty. We devised a new task testing perceptual curiosity—the desire for information about sensory stimuli (Berlyne, 1954; Nicki, 1970; Jepma et al., 2012)—and used it in conjunction with fMRI to understand how visual representations and uncertainty are related to curiosity.

Converging evidence shows that visual uncertainty is represented in a distributed network of structures that includes both primary sensory areas and higher-order associative areas. Evidence from monkey neurophysiology (Ma et al., 2006; Walker et al., 2023), human fMRI (van Bergen and Jehee, 2021; Geurts et al., 2022), and computational modeling (Meyniel et al., 2015) suggests that visual uncertainty is encoded alongside stimulus features in the primary visual cortex (V1). V1 orientation-tuned neurons represent the most likely orientation of a stimulus, and signal the uncertainty about that orientation in the distributed pattern of activity: a clear, high contrast stimulus evokes a focused activation of the neurons that are tuned to the stimulus, while an ambiguous, low contrast stimulus activates a broader population of neurons encoding a range of possible features. Thus, uncertainty is conveyed by the pattern of activity across a population of feature-tuned cells, a type of representation that we refer to as multivariate certainty (Russell and Reale, 2019).

Confidence ratings, in contrast, are encoded in higher-order structures, particularly in the frontal lobe (Del Cul et al., 2009; Fleming et al., 2010, 2014; Rounis et al., 2010; Lebreton et al., 2015; Gherman and Philiastides, 2018). Two key areas specifically involved in perceptual confidence are the ventromedial prefrontal cortex (vmPFC), which encodes visual confidence in multiple tasks (Lebreton et al., 2015; Hebscher et al., 2016; Gherman and Philiastides, 2018), and the anterior cingulate cortex (ACC), which also encodes visual confidence (Bang and Fleming, 2018; Geurts et al., 2022) and is implicated in cognitive control (Shenhav et al., 2013) and curiosity-driven information gathering (Jepma et al., 2012; Silvetti et al., 2023). These results are consistent with the broader role of the frontal lobe in metacognitive evaluation (Shimamura, 2000; Fleming et al., 2010), and with the fact that confidence—one's subjective sense of uncertainty—depends not only on sensory features but also behavioral factors like response heuristics, response bias, or context (Maniscalco et al., 2016; Peters et al., 2017; Zylberberg et al., 2012).

In contrast to the multivariate representation in the primary sensory cortex, frontal cortical areas convey a univariate representation of confidence—i.e., encode confidence through overall increases or decreases in their average BOLD response (Lebreton et al., 2015; Gherman and Philiastides, 2018). A prominent hypothesis is that these areas read out the population representations in sensory areas (Meyniel et al., 2015; Pouget et al., 2016), which transforms multivariate neural representations into simpler, lower-dimensional representations that can be more readily used for controlling behavior (DiCarlo and Cox, 2007; Russo et al., 2018). A recent study provides empirical evidence supporting this view, showing that the confidence of human observers is related to the multivariate certainty decoded from the BOLD response in V1 (Geurts et al., 2022).

A key open question, however, is whether and how this distributed circuitry relates to curiosity. The study of Geurts and colleagues used a traditional task in which participants were trained to discriminate the orientation of a Gabor stimulus and decoded multivariate certainty from human V1 based on the restrictive assumption that orientation is encoded by neurons with well-defined cosine tuning curves (Geurts et al., 2022; see also Walker et al., 2023; van Bergen and Jehee, 2021). In contrast, curiosity arises spontaneously and is evoked by complex natural stimuli that are represented in higher-level visual areas like the occipitotemporal cortex (OTC) that lack well-defined tuning curves. OTC is involved in visual recognition (Grill-Spector and Malach, 2004; Kar et al., 2019) and encodes biologically relevant stimulus categories using multivoxel activity patterns (Konkle and Oliva, 2012; Konkle and Caramazza, 2013; Long et al., 2018), but no study has quantified the certainty of its object representations.

We introduced three key innovations to understand whether and how the OTC visual representations are related to curiosity. First, we devised a new task in which participants reported their confidence and curiosity about ambiguous visual images rather than trivia questions. Second, we used synthetic images of animals and man-made objects ("texforms") that could be distorted according to a well-defined algorithm (Long et al., 2018) to activate OTC multivoxel category representations. Finally, we developed a trial-by-trial metric of OTC Certainty—the certainty conveyed by OTC multivoxel patterns—that made no assumptions about canonical tuning curves and was inspired instead by the modern machine learning literature (Hüllermeier and Waegeman, 2021).

Using these methods, we show that the OTC conveys a multivariate representation of visual uncertainty, which influences curiosity through a pathway mediated by the vmPFC. Consistent with findings from trivia questions (Gruber et al., 2014; Baranes et al., 2015), ratings of perceptual curiosity were inversely related with ratings of confidence about image identity. Also consistent with previous findings, BOLD activity in the vmPFC and ACC increased in function to confidence ratings and, reflecting the inverse relationship between confidence and curiosity, decreased with curiosity. Crucially, OTC Certainty about image category correlated with perceptual curiosity, and this link was mediated by the vmPFC but not the ACC, revealing a mechanism through which perceptual curiosity can arise from the sensory representation of a stimulus.

Materials and Methods

Participants

Thirty-two individuals (17 female; all right-handed; all normal or corrected-to-normal vision) participated for monetary compensation ($20/h; $40 in total). The participants' self-reported ages ranged between 18 and 35 (mean = 27.2; SD = 4.5), and their years of education ranged between 13 and 25 (mean = 16.1; standard deviation = 2.8). The study was approved by the Institutional Review Board at Columbia University. Participants were recruited via mailing lists and a research participation pool at Columbia University. Written informed consent was obtained from all participants. Participants also passed a health and safety screening on their eligibility for the MRI scanner.

Stimuli

To elicit curiosity, we used texforms (Long et al., 2018; Deza et al., 2019). We first collected 42 images of animals and 42 images of man-made objects from the Konkle Lab image database and normalized them for contrast and luminance across the whole set using the SHINE Toolbox in MATLAB. Then, we used an existing algorithm (Deza et al., 2019) that calculated thousands of first- and second-order image statistics from individual pooling regions overlaid across each image. Finally, we generated the texform by starting from a white noise display and coercing it to match the measured image statistics using stochastic gradient descent (100 iterations). The resulting texform looked like a distorted version of the original one. The size of the pooling regions (spatial pooling factor) determined the degree of distortion; all the images we used had a constant pooling factor of 0.28.

To investigate if low-level visual properties can predict our variables of interest (e.g., OTC Certainty), we measured the luminance, contrast, and spatial frequency of the texforms. Luminance was determined by calculating the mean of the intensity histogram, while root-mean-squared (RMS) contrast (i.e., standard deviation of the luminance distribution) was used to measure contrast, as is typical in vision studies using natural stimuli (Peli, 1990). Spatial frequency was calculated by first using 2-D Fast Fourier Transform (fft2 in Matlab), calculating power spectra in each dimension, and calculating the square root of the sum of squares of the resulting power spectra [i.e., $\mathrm{sqrt}(\mathrm{power}(x)^2 + \mathrm{power}(y)^2)$]. Average spatial frequency was then calculated as the slope of this resulting power spectrum and provided a measure of the amount of high versus low frequency information in the image. This approach to calculate spatial frequency content has been used in previous studies (Eskicioglu and Fisher, 1995; Li et al., 2001; Flitcroft et al., 2020).

Experimental design and statistical analysis

Design and procedure

Perceptual curiosity task. Stimuli were presented using the Psychophysics Toolbox for MATLAB (Psychtoolbox). For all trial components that required participants to enter a rating, participants did so on a continuous scale from 0 to 100 using an MR-compatible trackball. The initial slider position was randomized on every trial. Participants had up to 5 s to respond to each prompt; the trial advanced to the next screen after a response was entered.

The design of the perceptual curiosity task was inspired by Gruber et al. (2014) and Jepma et al. (2012). Participants were informed that they would view images that will be recognizable to different degrees and would be equally likely to depict and animal or man-made object and, on each of 84 trials, were shown a texform (see above, Stimuli) that was randomly drawn from either category. The texform remained on the screen for 4 s, and participants were instructed to come up with their best guess for what the original (undistorted) image was. Next, participants were prompted to rate their confidence in their best guess for the original image and their curiosity to see the original image. Finally, participants viewed the original image for 2 s. Trials were divided evenly into four runs. Importantly, participants were paid a fixed amount of $40 regardless of performance, ensuring that their confidence and curiosity ratings are independent of monetary incentives.

Localizer task. After this task, participants completed an unannounced localizer run, in which they viewed alternating miniblocks of clear (undistorted) animal and man-made object images that were not seen earlier in the task. Each of the 24 miniblocks (12 animal miniblocks and 12 man-made miniblocks) consisted of the presentation of 20 images presented in rapid succession (333 ms per image and 333 ms interstimulus interval). Between miniblocks, participants were presented with a fixation screen (13 s), which allowed for separation of BOLD activity between miniblocks. During image presentation, participants completed a one-back cover task, in which they were asked to detect and respond to repeat images using a button box.

Behavioral analysis

All mixed-effects models of behavior were conducted using the *fitlme* function in MATLAB.

Mixed-effects modeling of curiosity. To examine the relation between confidence and curiosity ratings, we constructed two mixed-effects models. In the quadratic model, curiosity was predicted by confidence, confidence^2, and participant-specific random slopes (confidence|participant and confidence^2|participant) and intercepts (1|participant):

$$\begin{aligned}\text{curiosity} \sim\ & \text{confidence} + \text{confidence}^2 \\ & + (1 + \text{confidence} + \text{confidence}^2 | \text{participant}).\end{aligned} \quad (1)$$

In the linear model, curiosity was predicted by confidence and participant-specific random slopes (confidence|participant) and intercepts (1|participant):

$$\text{curiosity} \sim \text{confidence} + (1 + \text{confidence} | \text{participant}). \quad (2)$$

To examine whether the relation between confidence and curiosity persists when including low-level visual properties as covariates, we constructed a mixed-effects model in which curiosity was predicted from confidence, luminance, contrast, and spatial frequency, as well as participant-specific random slopes (confidence|participant, luminance|participant, contrast|participant, and spatial frequency|participant) and intercepts (1|participant):

$$\begin{aligned}\text{curiosity} \sim\ & \text{confidence} + \text{luminance} + \text{contrast} \\ & + \text{spatial frequency} + (1 + \text{confidence}|\text{participant}) \\ & + (1 + \text{luminance}|\text{participant}) + (1 + \text{contrast}|\text{participant}) \\ & + (1 + \text{spatial frequency}|\text{participant}).\end{aligned} \quad (3)$$

Independent variables across both models were *z*-scored within participant.

MRI acquisition

MRI data were collected on a 3 T Siemens Magnetom Prisma scanner with a 64-channel head coil. Functional images were obtained with a multiband echoplanar image (EPI) sequence (repetition time, 2 s; echo time, 30 ms; flip angle, 80° acceleration factor, 3; voxel size, 2 mm isotropic; phase encoding direction: posterior to anterior), with 69 axial slices (14° transverse to coronal) acquired in an interleaved fashion. There were five functional runs: four for the perceptual curiosity task and one for the localizer task. Whole-brain high-resolution (1.0 mm isotropic) T1-weighted structural images were acquired with a magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence.

fMRI analysis

Software. Preprocessing and analyses were performed using FEAT, FNIRT, and command-line functions in FSL (e.g., fslmaths). Subsequent analyses were performed using custom MATLAB scripts. Code is available upon request.

ROI definition. The vmPFC region of interest (ROI) was based on Mackey and Petrides (2014), but voxels were removed that overlapped with the corpus callosum. The OTC ROI and ACC ROI were derived

from the Harvard-Oxford Brain Atlas using the atlas tool (threshold, 50) in fslview. All ROIs were in 1.0 mm MNI space.

Preprocessing. We performed brain extraction (using BET), motion correction (using MCFLIRT in FEAT), high-pass filtering (cutoff, 100 ms), and spatial smoothing (3 mm FWHM Gaussian kernel). Functional images were registered to the standard 1 mm MNI152 structural image using a nonlinear transformation with 12 degrees of freedom. This registration was done on first-level analysis output (beta maps from individual trials or miniblocks).

GLM #1: localizer GLM. We implemented a GLM for each participant that predicted BOLD activity of every voxel from a design matrix that modeled each localizer miniblock as a boxcar from the onset of the first image in each miniblock to the offset of the last image in each miniblock (16.66 s). Thus, for each of the 24 miniblocks, we included one regressor in the design matrix, yielding one beta map for each miniblock. This design matrix also contained fixed-body motion-realignment regressors (x, y, z, pitch, roll, and yaw) and their respective first derivatives. All regressors were convolved with a double-gamma hemodynamic response function. Autocorrelations in the time series were corrected with FILM prewhitening.

Formation of animal and man-made object templates from GLM#1. To create the animal template, beta maps for the 12 animal miniblock regressors (from GLM #1) were averaged within the OTC ROI. To create the man-made object template, beta maps for the 12 man-made object miniblock regressors (from GLM #1) were averaged within the OTC ROI.

GLM #2: single trial GLM. We implemented a GLM for each participant that predicted BOLD activity of every voxel from a design matrix that separately modeled the texform presentation on each trial as a 4 s boxcar (i.e., 21 texform regressors per run). This design matrix also contained fixed-body motion-realignment regressors (x, y, z, pitch, roll, and yaw) and their respective first derivatives. Other regressors modeled trial components that were not of primary interest for fMRI analyses: one regressor that modeled all the confidence ratings periods, each as a boxcar with a duration equal to the participant's response time (RT); one regressor that modeled all the curiosity ratings, each as a boxcar with a duration equal to the participant's RT; and one regressor that modeled all the clear image presentations, each as a boxcar with a duration of 2 s. Each run was modeled separately, resulting in four different models per participant. All regressors were convolved with a double-gamma hemodynamic response function. Autocorrelations in the time series were corrected with FILM prewhitening.

Univariate activity in vmPFC, ACC, and OTC during the texform period was obtained by averaging beta values across voxels, separately for each texform presentation. Multivariate activity patterns in OTC during the texform period were obtained by extracting the activity pattern across voxels, separately for each texform presentation.

Quantification of OTC certainty. We calculated r_a, Pearson's correlation coefficient between the animal template and texform pattern on a given trial, and r_{mm}, Pearson's correlation coefficient between the man-made template and texform pattern. We quantified OTC Certainty on every trial as the product of two terms: (1) relative evidence, the absolute value of the difference between r_a and r_{mm} and (2) mean evidence, the mean of r_a and r_{mm}.

Mixed-effects modeling. We constructed mixed-effects models using the *fitlme* function in MATLAB to examine relationships between brain measures and between brain measures and behavior:

To examine the relationship between confidence and OTC Certainty, we constructed two mixed-effects models. In the quadratic model, confidence was predicted by OTC Certainty, OTC Certainty^2, and participant-specific random slopes and intercepts:

$$\begin{aligned}\text{confidence} \sim\ & \text{OTC Certainty} + \text{OTC Certainty}^2 \\ & + (1 + \text{OTC Certainty} + \text{OTC Certainty}^2 | \text{participant}).\end{aligned} \quad (4)$$

In the linear model, confidence was predicted by OTC Certainty and participant-specific random slopes and intercepts:

$$\begin{aligned}\text{confidence} \sim\ & \text{OTC Certainty} \\ & + (1 + \text{OTC Certainty} | \text{participant}).\end{aligned} \quad (5)$$

To examine the relation between curiosity and OTC Certainty, we constructed two mixed-effects models. In the quadratic model, curiosity was predicted by OTC Certainty, OTC Certainty^2, and participant-specific random slopes and intercepts:

$$\begin{aligned}\text{curiosity} \sim\ & \text{OTC Certainty} + \text{OTC Certainty}^2 \\ & + (1 + \text{OTC Certainty} + \text{OTC Certainty}^2 | \text{participant}).\end{aligned} \quad (6)$$

In the linear model, curiosity was predicted by OTC Certainty and participant-specific random slopes and intercepts:

$$\text{curiosity} \sim \text{OTC Certainty} + (1 + \text{OTC Certainty} | \text{participant}). \quad (7)$$

Similar models were constructed to examine the relationships between confidence and vmPFC activity, curiosity and vmPFC activity, OTC Certainty and vmPFC activity, confidence and ACC activity, curiosity and ACC activity, and OTC Certainty and ACC activity.

To examine whether the relationship between OTC Certainty and curiosity persists when including low-level visual properties as covariates, we constructed a mixed-effects model in which curiosity was predicted from OTC Certainty, luminance, contrast, and spatial frequency, as well as participant-specific random slopes (curiosity|participant, luminance|participant, contrast|participant, and spatial frequency|participant) and intercepts (1|participant). Similar models were constructed to examine whether vmPFC/ACC activity could be predicted by low-level visual properties:

$$\begin{aligned}\text{curiosity} \sim\ & \text{OTC Certainty} + \text{luminance} + \text{contrast} \\ & + \text{spatial frequency} + (1 + \text{OTC Certainty} | \text{participant}) \\ & + (1 + \text{luminance} | \text{participant}) + (1 + \text{contrast} | \text{participant}) \\ & + (1 + \text{spatial frequency} | \text{participant}).\end{aligned} \quad (8)$$

Independent variables across all models were z-scored within participant.

Model comparison. We compared the relative goodness-of-fit of the quadratic versus linear mixed-effects models using the Bayesian information criterion (BIC). Using the conventions from Raftery (1995), a BIC difference of ≥2 indicates evidence for one model over another.

Mediation analysis. To examine the hypothesis that univariate activity in our confidence ROIs mediates the correlation between OTC Certainty and curiosity, we used the Baron and Kenny approach to mediation (Baron and Kenny, 1986) implemented in MATLAB by Wager et al. (2009). This approach compares c' (e.g., the effect size of the linear term of OTC Certainty on curiosity when controlling for confidence ROI activity) with c (e.g., the effect size of the linear term of OTC Certainty on curiosity alone). Statistical mediation occurs when c' is smaller in magnitude than c. We included quadratic terms in all pairwise comparisons because some had a significant quadratic relationship; but we only examined the change of magnitude of the linear terms when testing for mediation, consistent with established mediation approaches. Similar results were obtained if we only incorporated linear terms in the mediation analysis. We z-scored the three variables of interest (OTC certainty,

confidence ROI activity, and curiosity ratings) within participant and bootstrapped with replacement to obtain a distribution of c' and c values across 1,000 iterations. We then took the difference of the c and c' distributions and calculated the fraction of this new distribution that did not exceed 0. We then doubled this value to obtain our two-tailed p value.

Results

The perceptual curiosity task

Thirty-two participants (17 female) completed a task (Materials and Methods, Design and procedure, Perceptual curiosity task) that measured perceptual curiosity while undergoing fMRI scanning. Participants were informed that they would view a series of images of animals or man-made objects that were distorted to evoke visual uncertainty (Fig. 1*A*). The images were "texforms," synthetic images that preserve some texture and form information but are difficult to recognize at the exemplar level (Long et al., 2018) and reliably activate OTC multivoxel patterns (Long et al., 2018). On each trial, participants viewed a texform that was randomly drawn from one category, followed by a delay period during which they were instructed to imagine their best guess for what the original (undistorted) image was and by prompts to rate their confidence in their best guess and their curiosity to see the original image (Fig. 1*B*). After providing their ratings, participants were shown the original image. This procedure is similar to studies of epistemic curiosity (Kang et al., 2009; Gruber et al., 2014; Baranes et al., 2015) except that participants reported their curiosity and confidence about a visual image rather than a trivia question; thus, the texforms can be considered the "questions" and the undistorted image the "answers" in our task. Participants reported all the ratings by rotating an MRI-safe trackball to position a cursor on a 0–100 scale, with the initial cursor position randomized to control for motor confounds. Each participant completed 84 trials divided evenly into four runs. To ensure that the ratings were unbiased by instrumental incentives, participants received only a fixed compensation for completing the task ($40) but no payoffs or feedback based on the ratings they gave. Thus, the ratings strictly reflected subjective confidence and curiosity independently of external (objective) benchmarks.

Confidence and curiosity ratings share a negative, quadratic relationship

We found that perceptual curiosity and confidence ratings were negatively related (Materials and Methods, Behavioral analysis): curiosity tended to peak at relatively low confidence and declined as individuals became more confident in their answer (Fig. 1*C*). The relationship had a significant quadratic component. A mixed-effects model with linear and quadratic terms produced significant coefficients for both terms (Eq. 1; $\beta_{linear} = -13.46$, $p < 0.0001$, 95% CI = [−15.8 −11.0]; $\beta_{quadratic} = -5.60$, $p < 0.0001$, 95% CI = [−7.21 −3.99]) and quantitative comparisons showed that models containing both terms were superior to containing only a linear term ($BIC_{quadratic} - BIC_{linear} = -180$). Analyses of random effects coefficients produced no evidence for effects of age (Pearson's correlations; age vs linear terms, $r = -0.1$, $p = 0.9$; age vs quadratic terms, $r = 0.01$, $p = 0.5$), years of education (linear terms, $r = -0.06$, $p = 0.8$; quadratic terms, $r = 0.008$, $p = 0.9$), or sex (two-sample t tests; linear terms, $t = 0.001$, $p = 0.8$; quadratic terms, $t = 0.002$, $p = 0.1$), consistent with the null effects of these demographic descriptors in other curiosity tasks (Kobayashi et al., 2019; Rischall et al., 2023). Similarly, we found no evidence for significant effects of low-level image properties. Models that included luminance, contrast, and spatial frequency as covariates (Materials and Methods; Eq. 3) reproduced the linear–quadratic relationship between curiosity and confidence ($\beta_{linear} = -13.46$, $p < 0.0001$, 95% CI = [−15.8 −11.0]; $\beta_{quadratic} = -5.58$, $p < 0.0001$, 95% CI = [−7.19 −3.98]), and showed that these properties were not reliable predictors of either curiosity [luminance ($\beta = 0.70$, $p = 0.11$, 95% CI = [−0.17 1.57]); contrast ($\beta = -0.36$, $p = 0.41$, 95% CI = [−1.2 0.51]); spatial frequency ($\beta = -0.73$, $p = 0.16$,

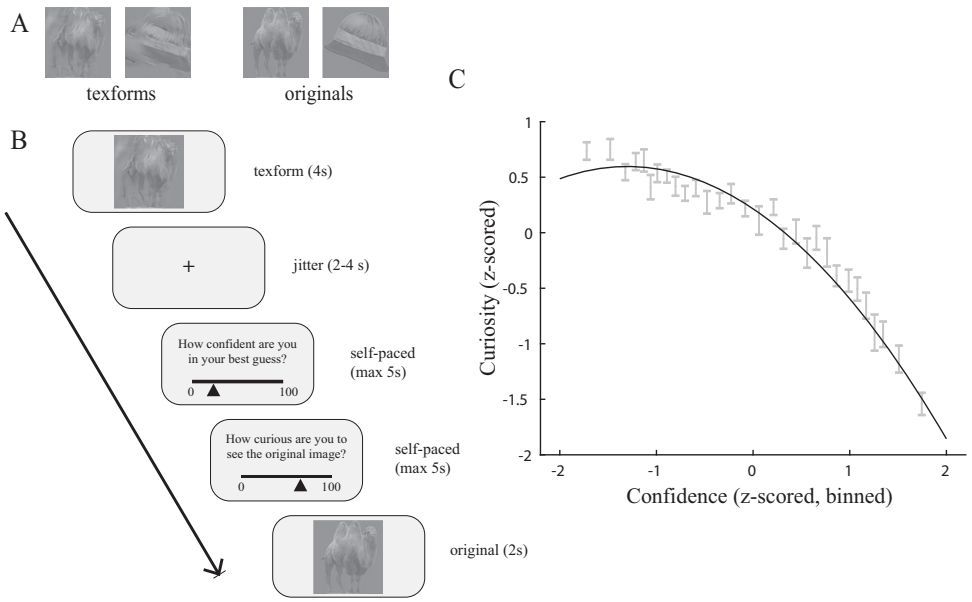

Figure 1. ***A***, Stimuli. Texforms of camel and hat (left), with corresponding clear images (right). Texforms are generated using texture synthesis algorithm from Deza et al. (2019) and preserve texture and form of original image while rendering image less recognizable. ***B***, Example trial. On each trial, participants view a texform for 4 s while simultaneously settling on a best guess for what the original image was. Participants then rate how confident they are in their best guess on a continuous scale from 0 to 100, rate how curious they are to see the original image from 0 to 100, and finally see the original image for 2 s. ***C***, Confidence predicts curiosity. All variables were z-scored within participant; the black curve is the curve of best fit (fixed effects) from the polynomial regression model (Materials and Methods; Eq. 1). For visualization purposes only, points were binned according to confidence (30 bins: equal number of points in each bin). The gray error bars indicate mean curiosity ±1 SEM.

95% CI = [−1.75 0.30])] or confidence [luminance ($\beta = -0.30$, $p = 0.60$, 95% CI = [−1.44 0.84]); contrast ($\beta = 1.20$, $p = 0.056$, 95% CI = [−0.07 2.3]); spatial frequency ($\beta = 0.76$, $p = 0.19$, 95% CI = [−0.38 1.90])]. In sum, consistent with epistemic curiosity (Kang et al., 2009; Gruber et al., 2014), perceptual curiosity and confidence showed a negative quadratic relationship, with curiosity being higher when participants had low or intermediate confidence in recognizing an item.

OTC certainty is positively correlated with confidence and negatively correlated with curiosity

To analyze the sensory underpinnings of perceptual curiosity, we focused on an anatomically defined ROI in the OTC, which encodes animal and man-made object categories across multivoxel activity patterns (Kriegeskorte et al., 2008; Konkle and Caramazza, 2013) and is thus a good candidate area for representing multivariate certainty about the categories in our task. Because existing measures of sensory certainty are limited to elementary visual features like orientation, our first goal was to develop a measure of sensory certainty from multivoxel activity patterns in OTC ("OTC Certainty").

To this end, we ran a localizer task in which participants viewed alternating miniblocks of undistorted animal and man-made object images (delivered unannounced after the main task and using images that differed from those in the main task; see Materials and Methods, Design and procedure, Localizer task). We then used a whole-brain general linear model to estimate the OTC multivoxel activity pattern evoked by each miniblock (Materials and Methods, fMRI analysis, GLM #1: localizer GLM). We verified that activity patterns evoked by miniblocks of the same category (i.e., animal/animal and man-made/man-made) showed higher pairwise correlations relative to patterns evoked by different-category miniblocks (i.e., animal/man-made; average Pearson's r, 0.80 vs 0.58; $p = 0.008$), confirming that our OTC ROI conveys reliable category representations. We then averaged the responses to miniblocks of each category to obtain, respectively, an "animal template" and a "man-made template"—the average multivoxel activity pattern expected for each image category.

Next, returning to the perceptual curiosity task, we measured the OTC activity pattern evoked by each texform (Materials and Methods, fMRI analysis, GLM #2: single trial GLM) and calculated Pearson's correlation coefficients between this activity pattern and each of our templates. For each texform-evoked activity pattern, we thus obtained two correlation coefficients—r_a, which measures its correlation with the animal template, and r_{mm}, which measures its correlation with the man-made template (Fig. 2*B*). We ascertained that texform responses were more highly correlated with the matching relative to non-matching template (average Pearson's r, 0.50 vs 0.43; paired t test $p = 0.01$), confirming that the OTC responses generalized across the perceptual curiosity and localizer tasks and the coefficients r_a and r_{mm} were valid measures of the similarity between a texform-evoked pattern and the multivariate representation of the animal and man-made categories.

Next, we defined a metric of OTC Certainty that combined two functions of r_a and r_{mm}—the mean of these terms and their absolute difference:

$$\text{OTC Certainty} = \text{mean}(r_a, r_{mm}) * |r_a - r_{mm}|.$$

This metric is motivated by the machine learning literature that emphasizes the importance of two types of uncertainty: model uncertainty and approximation uncertainty (Hüllermeier and Weigeman, 2021; see Discussion). The first term in our function, the mean of r_a and r_{mm}, corresponds to model uncertainty—the degree to which a measured response is consistent with the hypotheses that are considered by the model (i.e., correlates with one or both templates). The second term in the function, the absolute difference between r_a and r_{mm}, corresponds to approximation uncertainty—the degree to which a measured response distinguishes between the alternative hypotheses (i.e., correlates better with one template vs the other).

To better illustrate this new metric, Figure 2*C* represents the OTC texform responses into two ways: as a function of the mean and absolute differences of r_{mm} and r_a (left) and as a function of the raw values of r_{mm} and r_a (right). In each plot, the x- and y-coordinates of each point correspond to one texform response and the color of the point indicates OTC Certainty.

As shown in the left plot, OTC Certainty increases (more yellow) along the positive diagonal—i.e., if both the mean and the absolute differences are high, reflecting the multiplicative combination of the terms. In the right plot, OTC Certainty increases as points move further away from the equality diagonal and as they move toward the upper and right portions of the plot. The former effect stems from the absolute difference term, which drives OTC Certainty higher as r_{mm} and r_a become less similar to each other (their absolute difference increases). The increase in the upper right quadrant stems from the mean (r_{mm}, r_a) term, which drives OTC Certainty higher as the coefficients become larger.

To further unpack these relationships, consider four individual texforms (labeled 1–4 in both panels) illustrating different ways in which a response can gain high or low OTC Certainty. Texforms 1 and 2 have high OTC Certainty, indicated by their yellow hue. In both plots, these texforms lie in the upper right quadrants indicating that they have high r_{mm} and r_a [and consequently, high mean (r_{mm}, r_a)]. The left plot shows that the texforms also have high absolute difference values, and the right plot shows that this results from the fact that Texform 1 is more consistent with the man-made versus animal template while Texform 2 is more consistent with the animal versus man-made template. These texform responses thus have both high model certainty (as they are captured by at least one of the templates in our model) and high approximation certainty (correlate better with one of these templates than the other), resulting in high OTC Certainty. Texforms 3 and 4, on the other hand, are ascribed low OTC Certainty through distinct mechanisms. Texform 3 has a low absolute difference between r_{mm} and r_a (left) reflecting the fact that its two coefficients, while high, are similar to each other (right). Texform 4 has a low mean of r_{mm} and r_a (left) reflecting the fact that its two coefficients, despite having a large absolute difference, are overall low (right). Thus, low OTC Certainty can independently result from model uncertainty or approximation uncertainty (i.e., the mean or the absolute difference terms), reflecting the fact that these terms are minimally correlated (left panel; $r = -0.1$; $p = 0.04$).

Having established the properties of the OTC Certainty metric, we next investigated whether and how it was related to confidence and curiosity ratings. We found that OTC Certainty had a positive relation with confidence ratings ($\beta = 1.95$; $p = 0.0008$; 95% CI = [0.80 3.09]; Eq. 4; Fig. 2*D*, top). Consistent with the negative relationship between confidence and curiosity, OTC Certainty also had a negative relation with curiosity ratings ($\beta = -1.21$; $p = 0.007$; 95% CI = [−2.08 −0.33]; Eq. 6; Fig. 2*D*, bottom). Both relationships were linear, with model comparisons favoring models of confidence and curiosity that contained only linear terms for OTC Certainty over those that contained

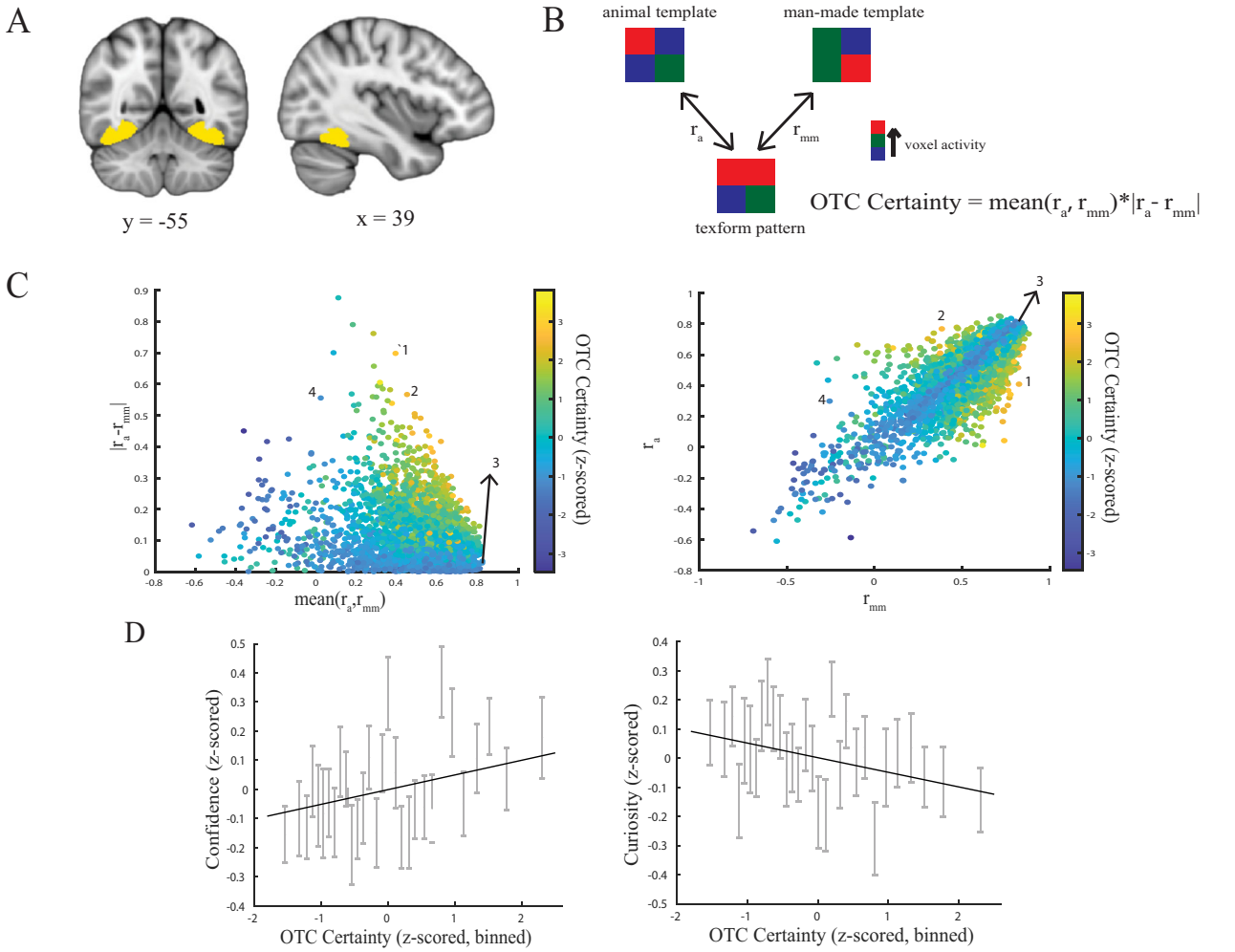

Figure 2. ***A***, Anatomical OTC ROI. Coronal (left) and sagittal (right) views showing the OTC ROI used for subsequent analyses (yellow). ***B***, Method for quantifying OTC Certainty. We extracted category templates from an independent localizer by averaging multivoxel patterns from animal and man-made miniblocks. We then correlated each category template with the multivoxel pattern elicited by the texform on each trial and calculated OTC Certainty as the product of the average and absolute value of the difference in these two correlations (see text for details). ***C***, Explanation of the OTC Certainty metric. The OTC Certainty metric (color) is shown as a function of the average and absolute differences of r_{mm} and r_a (left panel) and as a function of the r_{mm} and r_a values themselves (right). Each point shows the response to one texform in one participant. OTC Certainty values are *z*-scored within participants before pooling. Labels 1–4 indicate 4 example texforms (corresponding in the two panels) as described in the text. ***D***, OTC Certainty predicts confidence and curiosity. All variables were *z*-scored within participant; the black line is the line of best fit (fixed effects) from the linear regression model (Materials and Methods; Eqs. 4, 6). For visualization purposes only, points were binned according to OTC Certainty (30 bins: equal number of points in each bin). The gray error bars indicate mean confidence (top)/curiosity(bottom) and ±1 SEM in each OTC Certainty bin.

both linear and quadratic terms (confidence: $BIC_{quadratic}$–$BIC_{linear} = 17$; curiosity: $BIC_{quadratic}$–$BIC_{linear} = 23$). Thus, texforms that elicited higher OTC Certainty were more confidently recognized and evoked lower curiosity, while those that elicited lower OTC Certainty evoked higher curiosity.

Control analyses ruled out several confounds that may have affected these findings. One concern relates to pooling artifacts —the possibility that different participants contributed to different portions of the distributions in Figure 2*C*,*D*. However, in the above analyses, OTC Certainty and behavioral ratings were first *z*-scored within each participant before pooling across participants, factoring out individual differences. Moreover, we verified the results we robust to alternative normalizations that *z*-scored the raw r_{mm} and r_a or the mean and absolute difference terms rather than the final OTC Certainty metric. Second, a few of our texforms had r_{mm} or r_a values that happened to have negative signs (not significantly different from 0), but these constituted fewer than 4% of the data, and removing them from the analyses did not alter the final results. Finally, the findings in Figure 2*D* could not be explained by confounds related to head movements (which were included as nuisance regressors in GLM #2; Materials and Methods), cursor displacement (starting position was randomized on each trial), or the scaling factor controlling texform distortion (which was constant; Materials and Methods).

An important question is whether, in addition to OTC Certainty, curiosity may also relate to univariate OTC activity —but we found no evidence supporting this possibility. Although texforms evoked a wide range of average responses in the OTC ROI (a range of [−5.5, 3.5] *z*-score units vs [−3.5, 2.5] for curiosity ratings), a model relating curiosity to OTC univariate activity produced a nonsignificant coefficient ($\beta = -0.23$; $p = 0.63$; 95% CI = [−1.17 0.72]). Thus, curiosity is specifically associated with a multivariate pattern indicating OTC Certainty rather than merely the strength of the OTC response to a texform.

A second question is whether the associations between OTC Certainty and curiosity reflect low-level image properties. Arguing against this hypothesis, when we entered the luminance,

contrast, and spatial frequency of our texforms as covariates in the model linking OTC Certainty and curiosity (Materials and Methods), we obtained no significant coefficients for these properties [luminance ($\beta = 0.68$, $p = 0.12$, 95% CI = [−1.07 0.68]); contrast ($\beta = -0.19$, $p = 0.66$, 95% CI = [−1.07 0.68]); spatial frequency ($\beta = -0.71$, $p = 0.17$, 95% CI = [−1.72 0.31])], but replicated a significant effect of curiosity ($\beta = -1.15$; $p = 0.011$; 95% CI = [−2.03 −0.26]). Thus, the relationship between OTC Certainty and curiosity in our task was not contaminated by variation in the overall luminance, contrast, or spatial frequencies of the texforms.

As a different approach to this question, we asked if our OTC Certainty metric may have been explained by activity in the primary visual cortex (V1). Contrary to this hypothesis, we found no evidence that V1 encodes animal versus man-made categories. Specifically, we applied our analyses to V1 but first measuring the multivoxel activity pattern templates for animal and man-made objects in this area and then calculating Pearson's correlation coefficients between these templates and the texform-evoked activity patterns. In contrast with the OTC, texform-evoked activity patterns in V1 were not more correlated with the template of the true versus the alternate category (average Pearson's r, 0.57 vs 0.57; paired t test $p = 0.90$), and certainty derived from these patterns did not predict confidence ($\beta = 0.12$; $p = 0.8$; 95% CI = [−1.04. 1.30]) or curiosity ($\beta = 0.04$; $p = 0.9$; 95% CI = [−0.90, 0.99]). Since a category differentiation was necessary for our neural certainty metric, this makes it very unlikely that V1 activity accounts for our findings on OTC Certainty. Thus, OTC Certainty contributes to perceptual curiosity independently of lower-level representations (although the latter may contribute through distinct mechanisms, see Discussion).

vmPFC, but not ACC, mediates the link between OTC and curiosity

Our finding that OTC Certainty was related to both curiosity and confidence (Fig. 2), together with the strong association between the two ratings (Fig. 1C), raise the question of the nature of this relationship. One scenario, suggested by previous studies of sensory certainty (van Bergen and Jehee, 2021; Geurts et al., 2022), is that multivariate OTC Certainty is transformed into a univariate confidence representation elsewhere in the brain, which is in turn linked to curiosity. An alternative possibility, however, is that OTC Certainty influences curiosity ratings independently of univariate confidence representations. To adjudicate between these possibilities, we analyzed whether the relationship between OTC Certainty and curiosity may be mediated by two frontal brain areas implicated in perceptual confidence: the vmPFC (Fig. 3A; Mackey & Petrides Atlas) and the ACC (Fig. 3B; Harvard-Oxford Atlas).

We first verified that both areas showed univariate activity that scaled with confidence (Materials and Methods) as reported by previous studies (Lebreton et al., 2015; Bang and Fleming, 2018; Gherman and Philiastides, 2018). Indeed, confidence ratings were positively associated with univariate activity in the vmPFC ($\beta = 3.32$; $p < 0.0001$; 95% CI = [1.80 4.85]; Fig. 3A, left) and the ACC ($\beta = 2.34$; $p < 0.0001$; 95% CI = [0.88 3.79]; Fig. 3B, left). Both relationships were better fit by models with only a linear term relative to those with an additional quadratic term (vmPFC: $BIC_{quadratic} - BIC_{linear} = 23$; ACC: $BIC_{quadratic} - BIC_{linear} = 26$).

Consistent with the fact that confidence was inversely related to curiosity, univariate activity was also negatively associated with curiosity ratings, in vmPFC ($\beta_{linear} = -2.43$; $p < 0.0001$; 95% CI = [−3.58 −1.28]; Fig. 3A, right) and ACC ($\beta_{linear} = -1.14$; $p < 0.0001$; 95% CI = [−2.26 −0.02]; Fig. 3B, right). In the ACC, the relation between univariate activity and curiosity was equally well fit by models with and without a quadratic term ($BIC_{quadratic} - BIC_{linear} = 0$; $\beta_{quadratic} = -0.53$; $p = 0.005$; 95% CI = [−1.03 −0.02]) and in the vmPFC, the model that included a quadratic term produced a marginally better fit ($BIC_{quadratic} - BIC_{linear} = -3$; $\beta_{quadratic} = -0.66$; $p = 0.005$; 95% CI = [−1.13 −0.20]). Including low-level visual properties in the model revealed no significant relationships (ACC luminance $\beta = 1.34$, $p = 0.24$, 95% CI = [−2.56 1.37], contrast $\beta = -2.07$, $p = 0.06$, 95% CI = [−4.29 0.15], spatial frequency $\beta = -2.00$, $p = 0.11$, 95% CI = [−4.45 0.44]; vmPFC luminance $\beta = -0.61$, $p = 0.52$, 95% CI = [−2.49 1.27], contrast $\beta = -0.90$, $p = 0.35$, 95% CI = [−2.77 0.97], spatial frequency $\beta = -2.05$, $p = 0.08$, 95% CI = [−4.40 0.28]) while preserving the significant relationships with confidence and curiosity (confidence: ACC $\beta = 2.42$, $p = 0.0011$, 95% CI = [0.96 3.88]; vmPFC $\beta = 3.37$, $p < 0.0001$, 95% CI = [1.84 4.90]; curiosity: ACC $\beta_{linear} = -1.41$, $p = 0.017$, 95% CI = [−2.58 −0.25]; $\beta_{quadratic} = -0.54$, $p = 0.03$, 95% CI = [−1.04 −0.03]; vmPFC $\beta_{linear} = -2.48$, $p < 0.0001$, 95% CI = [−3.63 −1.33]; $\beta_{quadratic} = -0.66$, $p = 0.005$, 95% CI = [−1.12 −0.20]). Thus, both vmPFC and ACC were robustly activated by confidence consistent with the previous literature and had negative associations with curiosity, consistent with the negative relationship of this rating to confidence.

To understand the relationship between OTC Certainty, curiosity, and vmPFC and ACC, we performed mediation analyses (Materials and Methods, fMRI methods, Mediation analysis). According to the Baron and Kenny Mediation algorithm, mediation occurs if the effect of OTC Certainty on curiosity is reduced after controlling for activity in a frontal ROI (Baron and Kenny, 1986). Specifically, the algorithm generates two parameters indicating the strength of the associations between OTC Certainty and curiosity. Parameter c quantifies the strength of the association before considering the frontal ROI and parameter c' quantifies this strength after accounting for the frontal ROI activity. Mediation occurs if c' is significantly smaller than c—that is, activity in the frontal ROI explains away a significant portion of the association between OTC Certainty and curiosity.

We first verified that our data satisfy the necessary conditions for the mediation analysis—namely, significant pairwise associations between all the nodes in the model. As we showed above, curiosity was significantly associated with OTC Certainty (Fig. 2D), with vmPFC activity (Fig. 3A) and with ACC activity (Fig. 3B). Moreover, OTC Certainty was significantly associated with vmPFC activity (Fig. 4A, left; $\beta_{linear} = 5.55$; $p < 0.0001$; 95% CI = [2.91, 8.10]; $BIC_{quadratic} - BIC_{linear} = 17$) and with ACC activity (Fig. 4A, right; $\beta_{linear} = 9.00$; $p < 0.0001$; 95% CI = [6.70, 11.3]; $BIC_{quadratic} - BIC_{linear} = 3$). In the mediation model, these associations are reflected in, respectively, parameters a, b, and c, which were highly significant for both the vmPFC and the ACC models (Fig. 4B).

However, analysis of parameter c' that quantifies the mediated relationship between OTC Certainty and curiosity produced evidence for mediation by the vmPFC but not ACC. Consistent with the findings in Figure 2D, OTC Certainty showed a highly significant negative association with curiosity before accounting for activity in the frontal ROIs (parameter $c = -0.046$; $p < 0.009$; bootstrapped; Fig. 4B). When vmPFC was entered as a mediator, this relationship became nonsignificant (Fig. 4B, left; $c' = 0.036$; bootstrapped $p = 0.08$). Comparisons of parameters c and c'

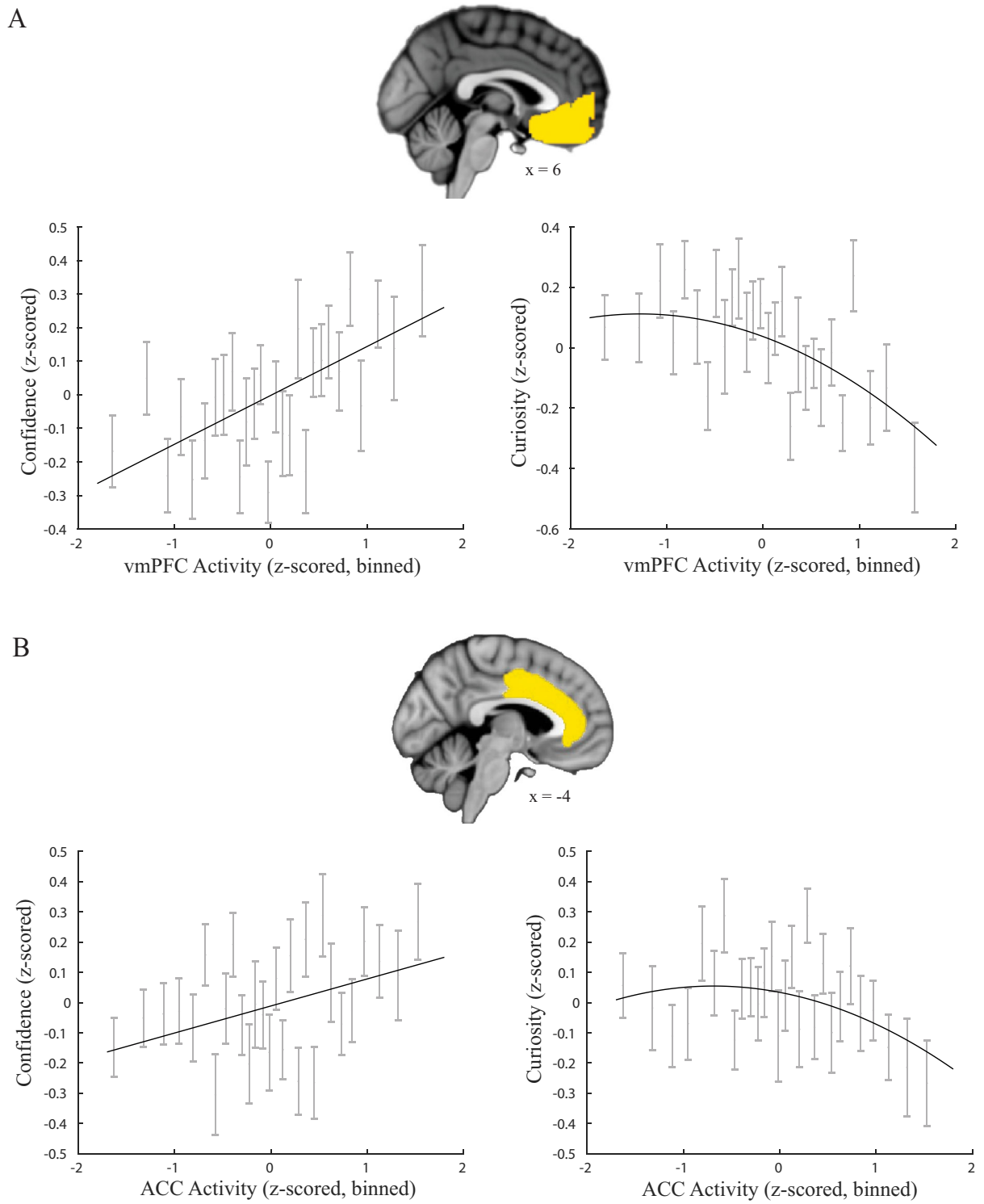

Figure 3. ***A***, The vmPFC encodes curiosity and confidence. The top panel shows the anatomical ROI for vmPFC (yellow). The bottom panels show the relationship between vmPFC activity and confidence and vmPFC activity and curiosity in the same format as in Figure 2*C*. ***B***, The ACC encodes curiosity and confidence. The top panel shows the anatomical ROI for the ACC (yellow). The bottom panels show the relationship between ACC activity and confidence and ACC activity and curiosity in the same format as Figure 3*A*.

showed that the latter was significantly reduced (closer to 0; bootstrapped c'-$c = 0.0109$; two-tailed $p < 0.001$), meaning that the vmPFC activity accounted for a significant portion of the association between OTC Certainty and curiosity. In contrast, when ACC was entered as a mediator, c' remained highly significant ($p < 0.001$; Fig. 4*B*, right), and it was statistically similar to c (bootstrapped c'-$c = 0.002$; two-tailed $p = 0.30$; Fig. 4*B*, right) and stronger than c' in the vmPFC model (Kolmogorov–Smirnov test on bootstrapped c' distributions; ks-stat = 0.60; $p < 0.001$).

Thus, the ACC was similar to the vmPFC in that it had a robust response to curiosity (Fig. 3*B*) and correlated with OTC Certainty (Fig. 4*A*, right) but differed specifically in providing no evidence that it mediated the association between OTC Certainty and curiosity. Together, the findings suggest that

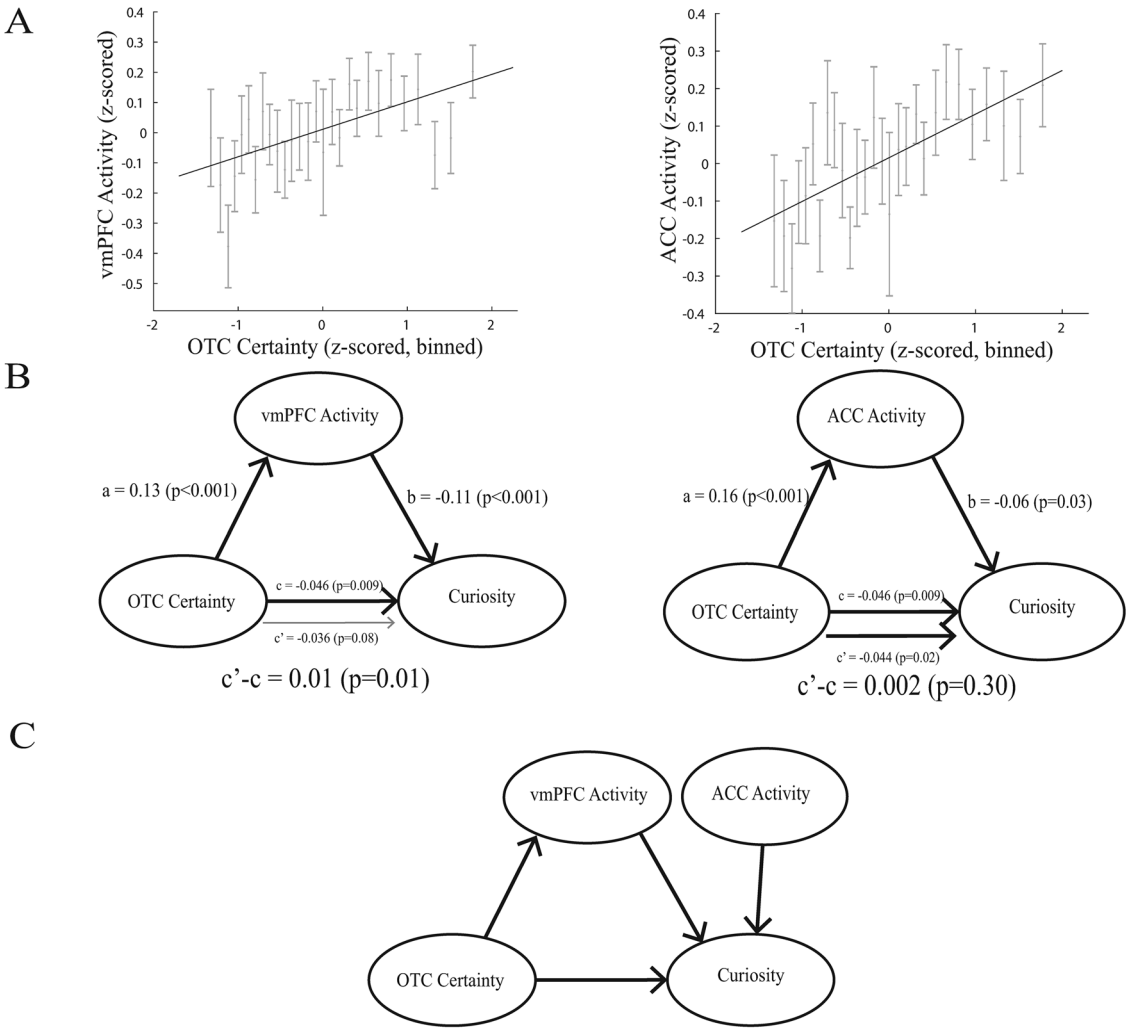

Figure 4. ***A***, OTC certainty significantly correlates with vmPFC and ACC activity, fulfilling a necessary condition for mediation analysis. Conventions as in Figure 2*D*. ***B***, Significant mediation occurs for the vmPFC (left) but not the ACC (right). In each path diagram, *a*, *b*, and *c* are the Pearson's correlation coefficients between, respectively, OTC Certainty and the frontal ROI, the frontal ROI and curiosity, and OTC Certainty and curiosity. *c*′ is the correlation coefficient between OTC Certainty and curiosity, when the frontal ROI response is controlled for. The numbers indicate the coefficients and their *p* values. Thick arrows show significant coefficients ($p < 0.05$). ***C***, Cartoon summarizing the functional interactions supported by the results. OTC Certainty contributes to curiosity through a pathway that is mediated by vmPFC, while the ACC contributes through a distinct mechanism. Note that the arrows in this diagram do not imply direct anatomical connections, but only functional associations detected through correlation or mediation analyses.

OTC Certainty is associated with curiosity through a mechanism that is mediated by the vmPFC, while the relation between ACC and curiosity emerges through a distinct mechanism.

Discussion

We tested the novel hypothesis that a neural representation of sensory certainty is transformed into a univariate confidence signal that generates curiosity. We used a new task in which participants rated their confidence and curiosity about images that were difficult to recognize and showed that perceptual curiosity had a negative relation with confidence analogous to findings on epistemic curiosity (Kang et al., 2009; Baranes et al., 2015). By using synthetic images known as "texforms" (Long et al., 2018), we elicited multivoxel representations of animate and inanimate categories in the OTC and derived a trial-by-trial measure of the sensory uncertainty these representations conveyed. We showed that OTC Certainty was correlated with curiosity ratings and this relationship was specifically mediated by the vmPFC but not the ACC. These results go beyond previous studies that merely identified brain regions encoding curiosity about trivia questions (Kang et al., 2009; Gruber et al., 2014) or presented blurry visual images without collecting curiosity and confidence ratings (Jepma et al., 2012) and suggest a mechanism by which a neural representation of the (un)certainty about an event can generate curiosity.

Empirical evidence that the visual cortex provides a multivariate representation of stimulus certainty comes from a study using imaging in human V1 during a task of orientation discrimination (Geurts et al., 2022). However, while that study decoded V1 uncertainty based on the assumption that V1 cells have cosine tuning for orientation (Walker et al., 2023; van Bergen and Jehee, 2021; Geurts et al., 2022), this assumption does not apply to higher-level visual regions like the OTC, where individual neurons and voxels show complex selectivity to stimulus categories (Kriegeskorte et al., 2008). Thus, our field lacks analytical models that can decode the uncertainty of higher-order representations of visual stimuli, semantic concepts, or categories.

Our new metric of OTC Certainty addresses this limitation using an analytical strategy that eschews restrictive assumptions

about tuning curves and relies instead on two types of uncertainty emphasized in the machine learning literature. Model certainty denotes the extent to which the data fall within the space of hypotheses that are considered by an analytical model; in contrast, approximation certainty denotes how well the model can differentiate between the hypotheses (Hüllermeier and Waegeman, 2021). We verified that our data satisfy two key assumptions of this strategy—that the correlation coefficients r_a and r_{mm} estimated the evidence in favor of the modeled categories and the OTC Certainty metric captures both types of uncertainty through the distinct, multiplicative contributions of two functional terms—the average and absolute differences of r_a and r_{mm} (Fig. 2*C*). Importantly, we found that the OTC Certainty for a texform correlated with the participants' confidence in recognizing the texform and with their curiosity in seeing the true image. These findings provide a new method for measuring perceptual certainty for visual categories, which may be in the future extended to other types of abstract information such as semantic concepts or categories.

One potential limitation of our approach is that multivariate certainty was measured at the level of animal and man-made object categories, while participants were instructed to generate best guesses about the distorted image with as much specificity as they could and may have generated guesses at the level of exemplars. However, our approach requires a single assumption: that generating a guess within one category activates patterns elicited by that category more than patterns elicited by other categories (regardless of the level of specificity). As we show, this assumption was met in our study, establishing the validity of our approach. Nevertheless, future studies can attempt to quantify multivariate certainty at the level of exemplars rather than categories and determine whether and how more granular representations account for the elicitation of curiosity.

Our analyses showed that ratings of confidence and curiosity were not influenced by low-level image properties (overall texform luminance, contrast, and spatial frequency) and that V1 responses lacked category specificity and could not account for OTC Certainty. However, given the evidence for a role of V1 in encoding uncertainty (Walker et al., 2023; van Bergen and Jehee, 2021; Geurts et al., 2022), it is possible that V1 did encode uncertainty about local visual features in ways we did not capture with our analysis and that such signals could be recruited to generate curiosity in different conditions—e.g., in a task in which participants were queried about local feature properties. Thus, our results leave open the possibility that different types of sensory uncertainty, encoded at different levels of the sensory hierarchy, contribute to perceptual confidence and curiosity through distinct mechanisms.

We showed that, in addition to their relationship with OTC Certainty, confidence and curiosity ratings correlated with average (univariate) responses in vmPFC and ACC. Although the vmPFC and ACC are sensitive to rewards (Kable and Glimcher, 2007; Shenhav et al., 2013) and curiosity is associated with intrinsic rewards (Litman, 2005; Kang et al., 2009; Gruber and Ranganath, 2019), the curiosity responses we found are likely separable from a subjective value response. While value responses have positive scaling—producing higher responses for higher rewards in both vmPFC and ACC (Kable and Glimcher, 2007; Cai and Padoa-Schioppa, 2012; Shenhav et al., 2013; Clithero and Rangel, 2014; Goh et al., 2021; Yee et al., 2021)—the curiosity response we report had negative scaling (lower responses for higher curiosity), inconsistent with a value response.

Our mediation analyses suggest that multiple mechanisms contribute to curiosity. One mechanism involves a mediated pathway whereby multivariate OTC Certainty is translated into a univariate confidence response in the vmPFC, which then correlates with curiosity. A second mechanism involves univariate activity in the ACC, which is associated with curiosity independently of OTC Certainty. Both mechanisms are defined in functional rather than anatomical terms and may be mediated by direct or indirect anatomical pathways.

Our finding that vmPFC mediates the link between OTC Certainty and curiosity is consistent with the reliable encoding of confidence in this area (Lebreton et al., 2015; Hebscher et al., 2016; Gherman and Philiastides, 2018) and with the direct anatomical connections between the vmPFC and the OTC (Catani et al., 2003; Furl, 2015). Interestingly, curiosity ratings had a linear relationship with OTC Certainty but a quadratic relationship with vmPFC activity—whereas in a mediated pathway, the two relationships may be expected to have a similar form (i.e., both should be linear or quadratic). One possibility is that we failed to detect a quadratic relationship between OTC Certainty and curiosity due to statistical noise (especially since the evidence for a quadratic relationship for vmPFC was moderate). Alternatively, an additional pathway may mediate the link between OTC Certainty and curiosity in which all the relationships are linear. Future work can examine whether these pathways exist and how they uniquely contribute to curiosity.

The fact that the ACC did not show significant mediation despite a robust encoding of confidence and curiosity suggests that it plays distinct roles in curiosity relative to vmPFC. One possibility is that the ACC computes confidence based on factors unrelated to OTC Certainty, such as response heuristics or biases (Zylberberg et al., 2012; Maniscalco et al., 2016; Peters et al., 2017) or even elementary visual features encoded in V1 (Geurts et al., 2022). Another, not mutually exclusive, possibility is that the ACC is not primarily involved in generating the subjective sensation of curiosity, but rather in recruiting cognitive functions to satisfy curiosity. This view is consistent with the roles of ACC in executive function (Shenhav et al., 2013) and attentional information gathering (Foley et al., 2017; Horan et al., 2019; White et al., 2018; Gottlieb et al., 2020; Silvetti et al., 2023) and suggests that the role of this area may be better revealed in tasks in which satisfying curiosity involves attentional effort (e.g., a participant is asked to find the information that relieves curiosity through effortful visual search). Thus, the specific role of the ACC in various types of decisions involving confidence and information demand remains an important topic for future investigations.

The evidence for significant mediation by the vmPFC implies a multistep mechanism, whereby OTC provides a multivariate representation of visual certainty, which is then translated into a univariate response to confidence in the vmPFC that triggers curiosity. What could be the computational advantage of this intricate process? We speculate that the answer is in providing different representations of uncertainty that are suited for different computational goals. A multivariate representation can convey (un)certainty while also signaling complex visual features. In contrast, a univariate representation is better suited for controlling actions—e.g., generating a lower-dimensional signal that regulates the level of arousal, curiosity, or attentional focus based on uncertainty. Thus, consistent with the representational untangling hypothesis (DiCarlo and Cox, 2007; Russo et al., 2018), we propose that the key advantage of a multistage process is to link

high-dimensional representations of sensory information to lower-dimensional representations that control actions.

Mechanistically, this process may be implemented through direct or indirect connections between the OTC and vmPFC. The OTC responses to visual categories have been well-documented and are believed to arise through a template-matching mechanism whereby a bottom-up stimulus activates a "template" reflecting category knowledge stored in the OTC (Mur et al., 2017). Our study, however, did not focus on how the category response emerges but on whether and how it is linked to curiosity. Our findings suggest that the templates that are activated by the texform in the OTC drive activity in vmPFC in a winner-takes-all fashion. Thus, the vmPFC acts like an exclusive-OR gate, generating high activity if one, and only one, of its inputs is active at a given time point but responding less if two inputs are weakly and/or similarly active. Thus, high vmPFC activity is associated with high confidence and with high model and approximation certainty in the OTC, while lower activity is associated with low confidence and low OTC Certainty. In principle, this readout may be implemented in the direct connections between the two areas (Catani et al., 2003; Furl, 2015). However, in natural behavior confidence monitoring requires high capacity and flexibility—i.e., individuals must monitor their confidence about multiple features that can rapidly change depending on the situation at hand. Thus, we believe it more likely that the readout involves indirect pathways perhaps including frontal and parietal areas, as can be determined in future research.

In conclusion, our results are consistent with the hypothesis that multivariate representations of certainty are transformed into univariate confidence in frontal cortex to generate curiosity. We speculate that this transformation might generalize outside of visual processing into other stimulus domains in which stimulus representations are probabilistic (Pouget et al., 2016; Lindskog et al., 2021) and likely described by multivariate certainty (Meyniel et al., 2015). Our findings shed light on the neural mechanisms of curiosity—specifically, how the brain may evaluate the neural representation of an event to generate high or low curiosity about that event.

Author's note

Mariam Aly is also affiliated with Department of Psychology, University of California Berkeley, Berkeley, CA 94720, USA.

References

Baranes A, et al. (2015) Eye movements reveal epistemic curiosity in human observers. Vision Res 117:81–90.
Bang D, Fleming SM (2018) Distinct encoding of decision confidence in human medial prefrontal cortex. Proc Natl Acad Sci USA 115:6082–6087.
Baron RM, Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 51:1173–1182.
Berlyne DE (1954) A theory of human curiosity. Br J Psychol 45:180–191.
Cai X, Padoa-Schioppa C (2012) Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex. J Neurosci 32:3791–3808.
Catani M, Jones DK, Donato R, Ffytche DH (2003) Occipitotemporal connections in the human brain. Brain 126:2093–2107.
Clithero JA, Rangel A (2014) Informatic parcellation of the network involved in the computation of subjective value. Soc Cogn Affect Neurosci 9:1289–1302.
Del Cul A, Dehaene S, Reyes P, Bravo E, Slachevsky A (2009) Causal role of prefrontal cortex in the threshold for access to consciousness. Brain 132:2531–2540.
Deza A, et al. (2019) Accelerated texforms: alternative methods for generating unrecognizable object images with preserved mid-level features. Comput Intell Neurosci. https://github.com/ArturoDeza/Fast-Texforms
DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341.
Eskicioglu AM, Fisher PS (1995) Image quality measures and their performance. IEEE Trans Commun 43:2959–2965.
Fleming SM, Weil RS, Nagy Z, Dolan RJ, Rees G (2010) Relating introspective accuracy to individual differences in brain structure. Science 329:1541–1543.
Fleming SM, Ryu J, Golfinos JG, Blackmon KE (2014) Domain-specific impairment of metacognitive accuracy following anterior prefrontal lesions. Brain 137:2811–2822.
Flitcroft D, Harb EN, Wildsoet CF (2020) The spatial frequency content of urban and indoor environments as a potential risk factor for myopia development. Invest Ophthalmol Vis Sci 61:42.
Foley NC, Kelly SP, Mhatre H, Lopes M, Gottlieb J (2017) Parietal neurons encode expected gains in instrumental information. Proc Natl Acad Sci U S A 114:E3315–E3323.
Furl N (2015) Structural and effective connectivity reveals potential network-based influences on category-sensitive visual areas. Front Hum Neurosci 9:253.
Geurts LS, Cooke JR, van Bergen RS, Jehee JF (2022) Subjective confidence reflects representation of Bayesian probability in cortex. Nat Hum Behav 6:294–305.
Gherman S, Philiastides MG (2018) Human VMPFC encodes early signatures of confidence in perceptual decisions. Elife 7:224337.
Goh AX, Bennett D, Bode S, Chong TT (2021) Neurocomputational mechanisms underlying the subjective value of information. Commun Biol 4:1346.
Golman R, Loewenstein G (2018) Information gaps: a theory of preferences regarding the presence and absence of information. Decision 5:143–164.
Gottlieb J, Cohanpour M, Li Y, Singletary N, Zabeh E (2020) Curiosity, information demand and attention priority. Curr Opin Behav Sci 35:83–91.
Gottlieb J, Oudeyer PY (2018) Towards a neuroscience of active sampling and curiosity. Nat Rev Neurosci 19:758–770.
Grill-Spector K, Malach R (2004) The human visual cortex. Annu Rev Neurosci 27:649–677.
Gruber MJ, Gelman BD, Ranganath C (2014) States of curiosity modulate hippocampus-dependent learning via the dopaminergic circuit. Neuron 84:486–496.
Gruber MJ, Ranganath C (2019) How curiosity enhances hippocampus-dependent memory: the Prediction, Appraisal, Curiosity, and Exploration (PACE) framework. Trends Cognitive Sci. 23:1014–1025.
Hebscher M, et al. (2016) Memory, decision-making, and the ventromedial prefrontal cortex (vmPFC): the roles of subcallosal and posterior orbitofrontal cortices in monitoring and control processes. Cereb Cortex 26:4590–4601.
Horan M, Daddaoua N, Gottlieb J (2019) Parietal neurons encode information sampling based on decision uncertainty. Nat Neurosci 22:1327–1335.
Hüllermeier E, Waegeman W (2021) Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Mach Learn 110:457–506.
Jepma M, Verdonschot RG, van Steenbergen H, Rombouts SA, Nieuwenhuis S (2012) Neural mechanisms underlying the induction and relief of perceptual curiosity. Front Behav Neurosci 6:5.
Kable JW, Glimcher PW (2007) The neural correlates of subjective value during intertemporal choice. Nat Neurosci 10:1625–1633.
Kang MJ, et al. (2009) The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. Psychol Sci 20:963–973.
Kar K, Kubilius J, Schmidt KM, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci 22:974–983.
Kidd C, Hayden BY (2015) The psychology and neuroscience of curiosity. Neuron 88:449–460.
Kobayashi K, et al. (2019) Diverse motives for human curiosity. Nat Hum Behav 3:587–595.
Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. J Neurosci 33:10235–10242.
Konkle T, Oliva A (2012) A real-world size organization of object responses in occipitotemporal cortex. Neuron 74:1114–1124.
Kriegeskorte N, et al. (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–1141.
Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015). Automatic integration of confidence in the brain valuation signal for Lebreton article. Nat Neurosci 18:1159–1167.

Li S, Kwok JT, Wang Y (2001) Combination of images with diverse focuses using the spatial frequency. Inf Fusion 2:169–176.
Lindskog M, Nyström P, Gredebäck G (2021) Can the brain build probability distributions? Front Psychol 12.
Litman JA (2005) Curiosity and the pleasures of learning: wanting and liking new information. Cogn Emot 19:793–814.
Loewenstein G (1994) The psychology of curiosity: a review and reinterpretation. Psychol Bull 116:75–98.
Long B, Yu CP, Konkle T (2018) Mid-level visual features explain the high-level categorical organization of the ventral stream. Proc Natl Acad Sci U S A 115:9015–9024.
Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432–1438.
Mackey S, Petrides M (2014) Architecture and morphology of the human ventromedial prefrontal cortex. Eur J Neurosci 40:2777–2796.
Maniscalco B, Peters MAK, Lau H (2016) Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. Atten Percept Psychophys 78:923–937.
Meyniel F, Sigman M, Mainen ZF (2015) Confidence as Bayesian probability: from neural origins to behavior. Neuron 88:78–92.
Mur M, et al. (2017) Functional readout analysis reveals nonlinear representational transformation from early visual to category-selective regions. J Vis 17:1230.
Murphy C, et al. (2021) Temporal proximity to the elicitation of curiosity is key for enhancing memory for incidental information. Learn Mem 28: 34–39.
Nicki RM (1970) The reinforcing effect of uncertainty reduction on a human operant. Can J Psychol 24:389–399.
Peli E (1990) Contrast in complex images. J Opt Soc Am A 7:2032–2040.
Peters M, et al. (2017) Perceptual confidence neglects decision-incongruent evidence in the brain. Nat Hum Behavior 1:0139.
Pouget A, Drugowitsch J, Kepecs A (2016) Confidence and certainty: distinct probabilistic quantities for different goals. Nat Neurosci 19:366–374.
Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111–163.
Rischall I, et al. (2023) Inefficient prioritization of task-relevant attributes during instrumental information demand. Nat Commun 14:3174.
Rounis E, Maniscalco B, Rothwell JC, Passingham R, Lau H (2010) Theta-burst transcranial magnetic stimulation to the prefrontal cortex impairs metacognitive visual awareness. Cogn Neurosci 1:165–175.
Russell R, Reale C (2019) Multivariate uncertainty in deep learning. Comput Sci 33:7937–7943.
Russo AA, et al. (2018) Motor cortex embeds muscle-like commands in an untangled population response. Neuron 97:953–966.e8.
Shapiro AD, Grafton ST (2020) Subjective value then confidence in human ventromedial prefrontal cortex. PLoS One 15:e0225617.
Shenhav A, Botvinick MM, Cohen JD (2013) The expected value of control: an integrative theory of anterior cingulate cortex function. Neuron 79:217–240.
Shimamura AP (2000) Toward a cognitive neuroscience of metacognition. Conscious Cogn 9:313–323.
Silvetti M, et al. (2023) A reinforcement meta-learning framework of executive function and information demand. Neural Netw 157:103–113.
van Bergen RS, Jehee JF (2021) TAFKAP: an improved method for probabilistic decoding of cortical activity. bioRxiv.
van Lieshout LLF, de Lange FP, Cools R (2020) Why so curious? Quantifying mechanisms of information-seeking. Curr Opin Behav Sci 35:112–117.
Wager TD, Waugh CE, Lindquist M, Noll DC, Fredrickson BL, Taylor SF (2009) Brain mediators of cardiovascular responses to social threat, part I: reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. Neuroimage 47:821–835.
Walker EY, Pohl S, Denison RN, Barack DL, Lee J, Block N, Ma WJ, Meyniel F (2023) Studying the neural representations of uncertainty. Nat Neurosci 26:1857–1867.
White MG, et al. (2018) Anterior cingulate cortex input to the claustrum is required for top-down action control. Cell Rep 22:84–95.
Yee DM, Crawford JL, Lamichhane B, Braver TS (2021) Dorsal anterior cingulate cortex encodes the integrated incentive motivational value of cognitive task performance. J Neurosci 41.
Zylberberg A, Barttfeld P, Sigman M (2012) The construction of confidence in a perceptual decision. Front Integr Neurosci 6:79.