Confirmation Bias through Selective Use of Evidence in Human Cortex

Hame Park[1,*], Ayelet Arazi[1,*], Bharath Chandra Talluri[1,2,*], Marco Celotto[3,4], Stefano Panzeri[3], Alan A Stocker[5] & Tobias H. Donner[1,6]

[1]Section Computational Cognitive Neuroscience, Department of Neurophysiology and Pathophysiology, University Medical Center Hamburg-Eppendorf, Martinistraße 52, Hamburg 20251, Germany

[2]Laboratory of Sensorimotor Research, National Eye Institute, National Institutes of Health, Bethesda, USA

[3]Institute for Neural Information Processing, Center for Molecular Neurobiology, University Medical Center Hamburg-Eppendorf, 20251, Hamburg, Germany

[4]Neural Computation Laboratory, Center for Neuroscience and Cognitive Systems, Istituto Italiano Di Tecnologia, 38068, Rovereto, Italy

[5]Department of Psychology, University of Pennsylvania, 3710 Hamilton walk Philadelphia, PA 19106 USA

[6]Bernstein Center for Computational Neuroscience Berlin, Humboldt-University Berlin, Philippstr. 13, Haus 6, 10115 Berlin

*Equal contribution authorship.

Keywords

Bounded rationality; evidence accumulation; information theory; encoding; decoding

ABSTRACT

Decision-makers often process new evidence selectively, depending on their current beliefs about the world. We asked whether such confirmation biases result from biases in the encoding of sensory evidence in the brain, or alternatively in the utilization of encoded evidence for behavior. Human participants estimated the source of a sequence of visual-spatial evidence samples while we measured cortical population activity with magnetoencephalography (MEG). Halfway through the sequence, participants were prompted to judge the more likely source category. Their processing of subsequent evidence depended on its consistency with the previously chosen category, but the encoding of evidence in cortical activity did not. Instead, the encoded evidence in parietal and primary visual cortex contributed less to the estimation report when that evidence was inconsistent with the previous choice. We conclude that confirmation bias originates from the way in which decision-makers utilize information encoded in the brain. This provides room for deliberative control.

INTRODUCTION

Adaptive behavior in an uncertain world often requires agents to continuously update their beliefs about the state of the environment based on sequences of noisy observations ("evidence")[1–5]. A prominent bias shaping this process is confirmation bias, the tendency to selectively gather and interpret new evidence that supports a pre-existing belief [6,7]. Confirmation bias affects important areas of human judgment and decision-making including medical diagnostics, judicial reasoning, or scientific hypothesis testing[7]. Remarkably, confirmation bias also manifests in elementary, value-neutral perceptual decisions under uncertainty, such as judging the direction of a weak visual motion signal embedded in noise: once participants committed to a categorical judgment about the source of a noisy evidence stream, subsequent evidence that is consistent with that initial judgment tends to have a stronger impact on a subsequent decision compared to inconsistent evidence[8]. These psychophysical observations open the door to identify the underlying mechanisms in the brain.

In current neurophysiological models of perceptual decision-making[9,10], sensory evidence is encoded in sensory cortex and then read out through a process including temporal accumulation in downstream brain areas[4,11–14]. In principle, selective interpretation of evidence could emerge from two different neural operations. First, the encoding of evidence in neural activity could be modulated such that its precision depends on the consistency of the evidence with a previous decision or belief. Second, the downstream readout of that evidence for a subsequent decision may be altered, such that belief-inconsistent evidence contributes less to the decision than belief-consistent evidence. Distinguishing between these two scenarios has important implications: a reduction in the precision of the encoding of inconsistent evidence implies unrecoverable information loss, while a consistency-dependent modulation of evidence readout is more malleable to strategic control. The latter mechanism could be counteracted, on the fly (within a given decision) or across longer timescales during learning.

Here, we investigated these scenarios by interrogating the encoding of sensory evidence in cortical population activity, as well as the readout of this evidence representation, during a well-defined context of selective evidence processing. We combined a novel behavioral task with the assessment of fine-grained patterns of human cortical population activity (through MEG source imaging[15]) and information-theoretic

approaches for a statistical dissection of evidence encoding and evidence readout[16,17]. The task was tailored to precisely quantify how individual evidence samples are encoded and processed after a categorical choice before reporting a final continuous decision.

RESULTS

During MEG recordings, 30 healthy participants performed a continuous visual decision-making task (Figure 1a). On each trial, they monitored a sequence of twelve evidence samples (checkered discs varying only in angular position) and finally reported their continuous estimate of the evidence source (i.e., mean of generative distribution, Figure 1b; mean squared error: 5.93°; Supplementary figure S1). In the key experimental condition ("Choice"), participants were prompted to judge the category of the evidence source (left versus right from a reference line) after the sixth sample (Figure 1a; mean accuracy: 81.4%, Figure 1d, black). Sorting the subsequent evidence samples by their consistency with the previous choice enabled us to identify its impact on the processing of subsequent evidence samples, at behavioral and neural levels.

Previous psychophysics work reported choice-induced biasing effects on subsequent evidence processing that were reminiscent of the known modulatory effects of selective attention[8], and biases in continuous decisions induced by categorical choices as well as visual cues[18]. We, therefore, included a control condition ("Cue"), without an intermittent choice, but instead a cue after six samples, which informed participants about the true source category (left versus right from reference) with 75% reliability (Figure 1a). Prompting an intermittent choice or presenting an intermittent cue of similar reliability allowed us to determine if any biases induced by these two categorical events would depend on whether the event was self-generated (Choice) or externally provided (Cue).

Selective weighting of sensory evidence in a continuous decision

To quantify the weighting of the sequentially presented evidence in the final decision, we computed the mutual information between the stimulus samples (S), for each sample position in the sequence, and the behavioral estimation report (E), an information-theoretic quantification of the weighting of S on final E commonly referred to as "psychophysical kernel"[4,12,19] (Figure S1c,d). In the following, we denote the resulting information-theoretic quantity as I(S;E). To minimize the impact of the systematic trial-by-trial variations of generative means on I(S;E) and to isolate the impact of stochastic fluctuations in the evidence on the final estimation decision[19], we only included trials whose generative means were near zero (Figure 1b; Methods). Simulations of non-biased observers showed that this procedure eliminated possible spurious effects originating from the correlations between the generative mean, cue or choice, and the final estimation, while ensuring enough data for computation of the information metrics (Methods). The same subset of trials was used in all comparisons of information measures between consistent and inconsistent samples reported below.

I(S;E) was significantly larger than zero for each sample, in both Choice and Cue conditions (Figure S1c,d), in line with the notion that subjects integrated information across the sample sequence in their final continuous decisions, as is commonly observed for categorical decisions[4,12,19]. There was no difference in I(S;E) between samples in the first interval that were consistent versus inconsistent with the subsequent

choice or cue (Figure 1f,g, s_1-s_6). Critically, in the second interval, I(S;E) was larger for consistent than for inconsistent samples (Figure 1f,g: black bars), for samples immediately following the cue/choice, in line with selective processing of sensory evidence induced by the preceding categorical choice or cue. This consistency effect on evidence weighting was stronger in the Choice compared to the Cue condition (Figure 1h). Within the subset of trials close to the category boundary analyzed here, the fraction of correct choices was slightly smaller than the fraction of correct (i.e., valid) cues (Cue:67.1% vs. Choice: 58.9%), when correctness was based on the generative mean (Figure 1c, Wilcoxon sign-rank test).

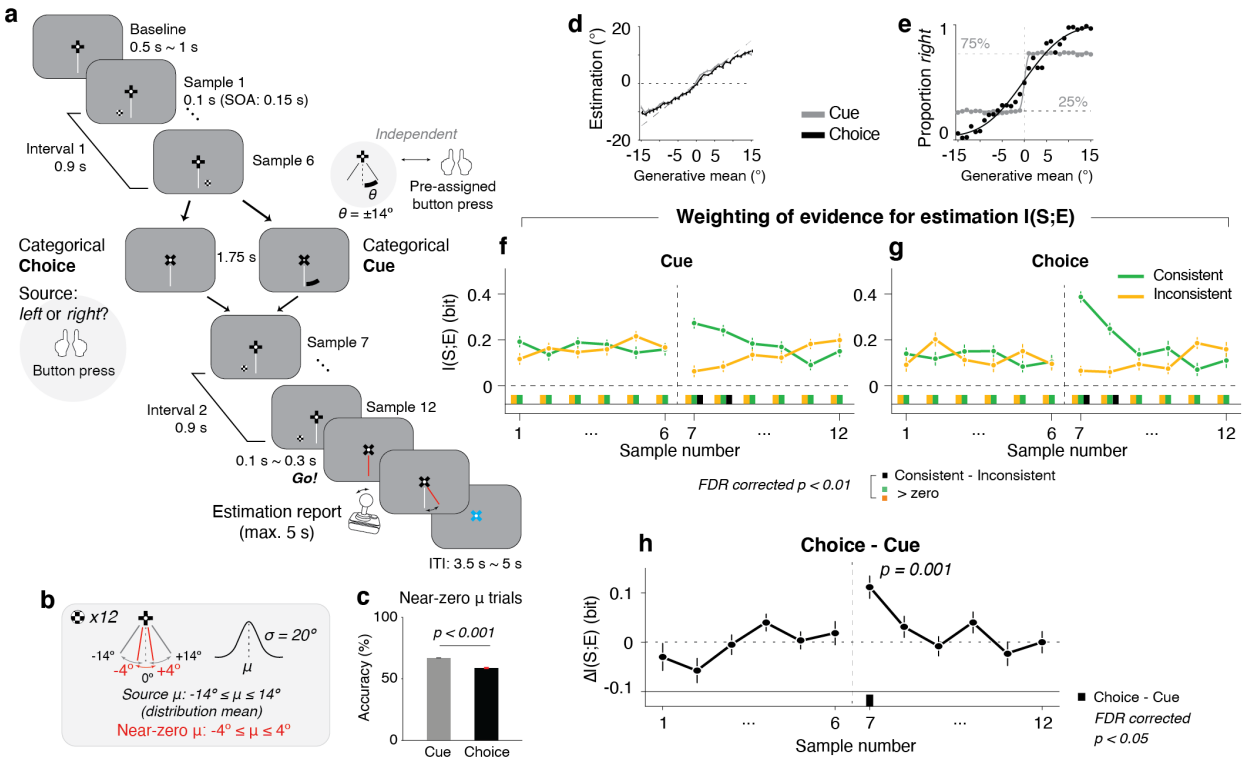

Figure 1. Task design and behavior.

(a) Schematic sequence of events in example trial. Visual evidence samples (polar angles from vertical reference line, of small discs at 5° of eccentricity) are drawn from Gaussian distributions (b). After the sequence, participants report their inference about the underlying source (distribution mean) by moving a joystick. After six samples, participants either report a categorical judgment about the source, or receive a categorical Cue (see main text). After Cue presentation, participants make a pre-assigned (left or right, blocked) button press, to match Choice and Cue conditions in terms of motor output. **(b)** Range of generative means (μ) and standard deviation (σ). Red font indicates the near-zero generative mean trials used for consistency effect analyses (-4° to 4°) **(c)** Accuracy (reliability) of the intermediate choice and cue for the near-zero trials (b). **(d)** Group average estimations as a function of generative mean for both conditions. **(e)** Proportion of "right" choices as a function of generative mean (Choice; first 6 samples). For comparison, the fraction of "right" Cues are plotted against the first 6 sample mean for Cue. Lines: fits of a sigmoid function on group data. Dots are group averages. **(f,g)** Impact of "sensory evidence" (samples) on estimation reports, computed as I(S;E), the mutual information between sample (S) and estimation report (E), for Cue (f) and Choice (g) conditions, for the entire stimulus sequence. I(S;E) is evaluated dependent on the consistency of each sample with the preceding (upcoming) choice or cue in the 2nd interval (1st interval). Vertical lines; standard error of mean across participants; thick lines; group average. Colored and black bars below 0, significant differences from 0 or between consistent and inconsistent, respectively (permutation tests, $p < 0.01$, FDR corrected). **(h)** Difference of consistency effects on I(S;E) between Choice and Cue conditions (consistent-inconsistent difference in f and g), tested with two-sided permutation test of mean (FDR corrected at $p = 0.05$). N=30 participants for all panels; dots, group mean; error bars, SEM.

These behavioral results are in line with recent psychophysics work on choice consistency-dependent evidence re-weighting in other tasks[8,18]. The results extend previous work by (i) delineating the precise time-course (across evidence samples) of the consistency effect and (ii) establishing such a consistency effect also for external categorical cues, and (iii) showing that the consistency effect is amplified when the bias-inducing event is self-generated (Choice) compared to when it is externally provided (Cue).

Encoding and readout of sensory evidence across cortex

We used the MEG data to track the encoding and readout of individual evidence samples. We reconstructed the patterns of sample-evoked neural population responses within well-defined cortical regions (Methods)[12] and quantified the information about task-relevant variables contained in these neural responses (Figure 2a). We performed comprehensive mapping of information measures across the cortical surface (Figure 2 b,d,f) complemented by focused analyses of "Dorsal Stream Visual Cortex"[20], a group of visual field maps[21,22] within and around the intraparietal sulcus (henceforth referred to as "dorsal visual cortex"; Methods) to illustrate the time courses of information measures (Figure 2c,e,g). Dorsal visual cortex has an established role in visual-spatial behavior[23–25]. We quantified three information-theoretic measures (whose relationship is schematized as a Venn diagram in Figure 2a, right): (i) I(S;R), the information about the sensory evidence sample S encoded by the neural population response R; (ii) I(R;E), the information about the upcoming behavioral (estimation) report E encoded by R; and, critically, (iii) the intersection information II(S;R;E), a statistical measure of how much of the sensory evidence sample information encoded in R is read out to inform the behavioral estimation report[17]. Figure 2 depicts those measures, as time courses following the onset of each sample (in dorsal visual cortex) and as maps across all areas (pooled across samples), for the trial interval of interest that followed the categorical choice or cue.

Our analyses yielded spatio-temporal profiles of information measures that were in line with the profiles obtained in (stimulus and decision) decoding approaches in similar tasks[4,12]. I(S;R) peaked ~150 ms in many visual cortical areas after sample onset (Figure 2c). I(R;E), the information about the upcoming estimation report (Figure 2d-e) was likewise present in many cortical areas, here in a more sustained fashion (Figure 2e), because it built up gradually throughout the processing of all evidence samples. In the Choice condition, I(R;E), was also prominent in frontal and parietal cortical regions involved in action planning (Figure 2d). Dorsal visual cortex exhibited robust I(S;R) and I(R;E) (Figure 2b-e), corroborating our focus on this brain region. The profile of intersection information (measure of readout of encoded evidence) II(S;R;E) exhibited spatio-temporal features resembling those of both I(S;R) and I(R;E) again with a robust effect in dorsal visual cortex (Figure 2f,g). Taken together, our results are in line with the idea that the sample information encoded in dorsal visual cortex was read out for behavioral estimation reports (Figure 2).

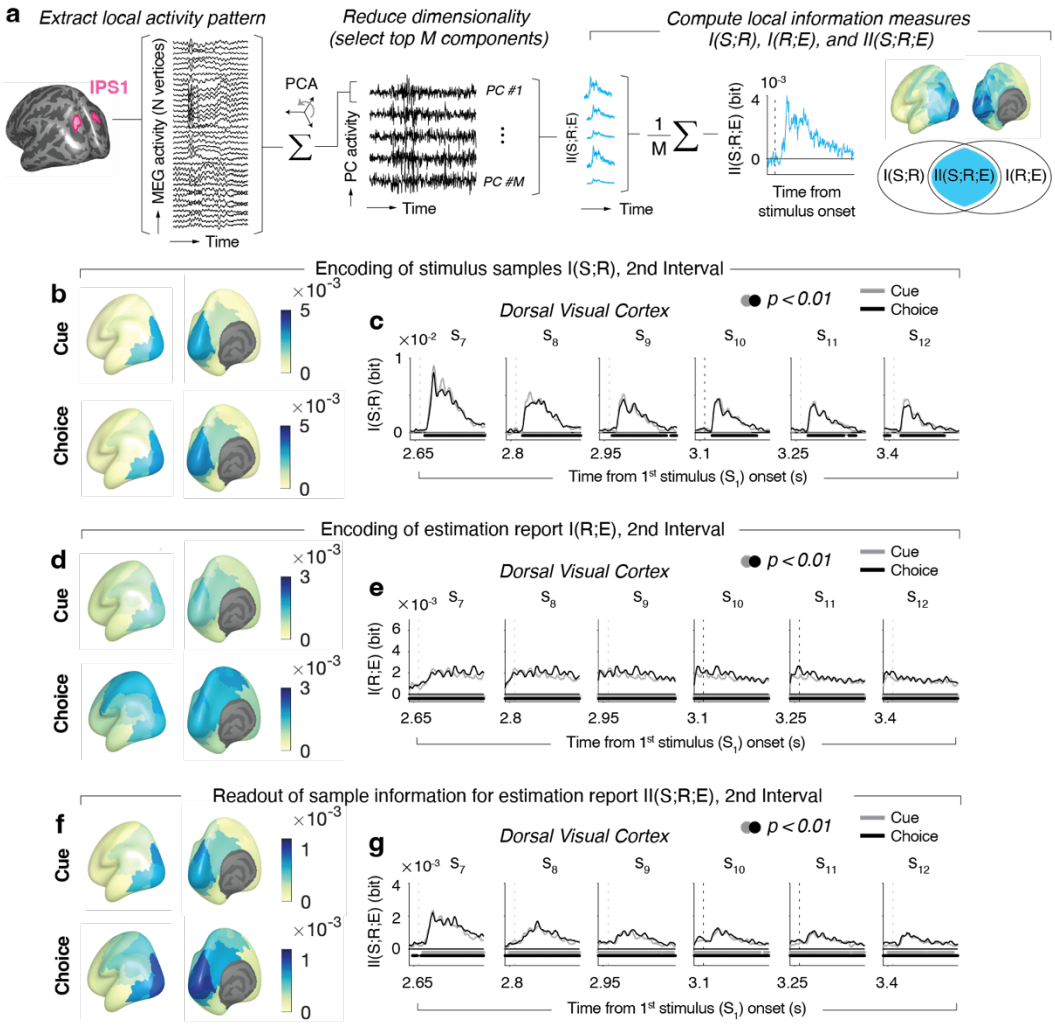

Figure 2. Neural encoding of evidence and estimation report in cortical population responses.
(a) Schematic of analysis approach. The dimensionality of source-reconstructed neural activity of all vertices per cortical area (example: IPS1 from dorsal visual cortex) was reduced with principal component analysis (PCA). *M* top components (between 1 and 15) that jointly explained 90 % of the variance were selected for all (information theoretic and decoding) analyses. Stimulus sample information I(S;R), estimation information I(R;E), and intersection information II(S;R;E) were computed for each component and then averaged (see main text for details). **(b)** Maps of group average I(S;R) following sample onset (averaged across samples from second trial interval, positions 7-12). **(c)** Time course of I(S;R) in dorsal visual cortex for each sample position of second trial interval. **(d-e)** As in (b-c), but for I(R;E). **(f-g)** As (b-c), but for II(S;R;E). For all maps, we averaged information measures across the interval 0.1 s – 0.5 s after sample onset, which contained most of the stimulus information (see panel c). Time courses were smoothed with a Gaussian kernel (sigma = 37.5 ms). Bars beneath time courses indicate time points significantly different from pre-trial (-0.5 s - 0 s baseline (two-sided non-parametric cluster-based test, $p < 0.01$) for Choice (black) or Cue (gray). N=30 participants for all panels.

Readout of stimulus information depends on consistency with previous choice or cue

In order to illuminate the neural basis of the effect of evidence consistency with the preceding choice on the final estimate, we compared the neural information measures between evidence samples that were consistent with the intermittent choice (or cue), and those which were inconsistent. Because we found behavioral consistency effects on I(S;E) for samples 7 and 8 right after the choice/cue (Figure 1f,g), we combined

these samples for our analyses of neural information measures to boost sensitivity (Figure 3; see Figure S2 for the corresponding maps for the matching samples 1 and 2 in the first trial interval and Figure S3 for non-thresholded maps across the complete trial). Consistency-dependent evidence encoding would predict higher I(S;R) for consistent than inconsistent samples. The data were inconsistent with this hypothesis: for both Cue and Choice conditions, there was no difference in I(S;R) between consistent and inconsistent samples, in dorsal visual cortex (Figure 3a,c), nor any other cortical area (Figure 3b,d).

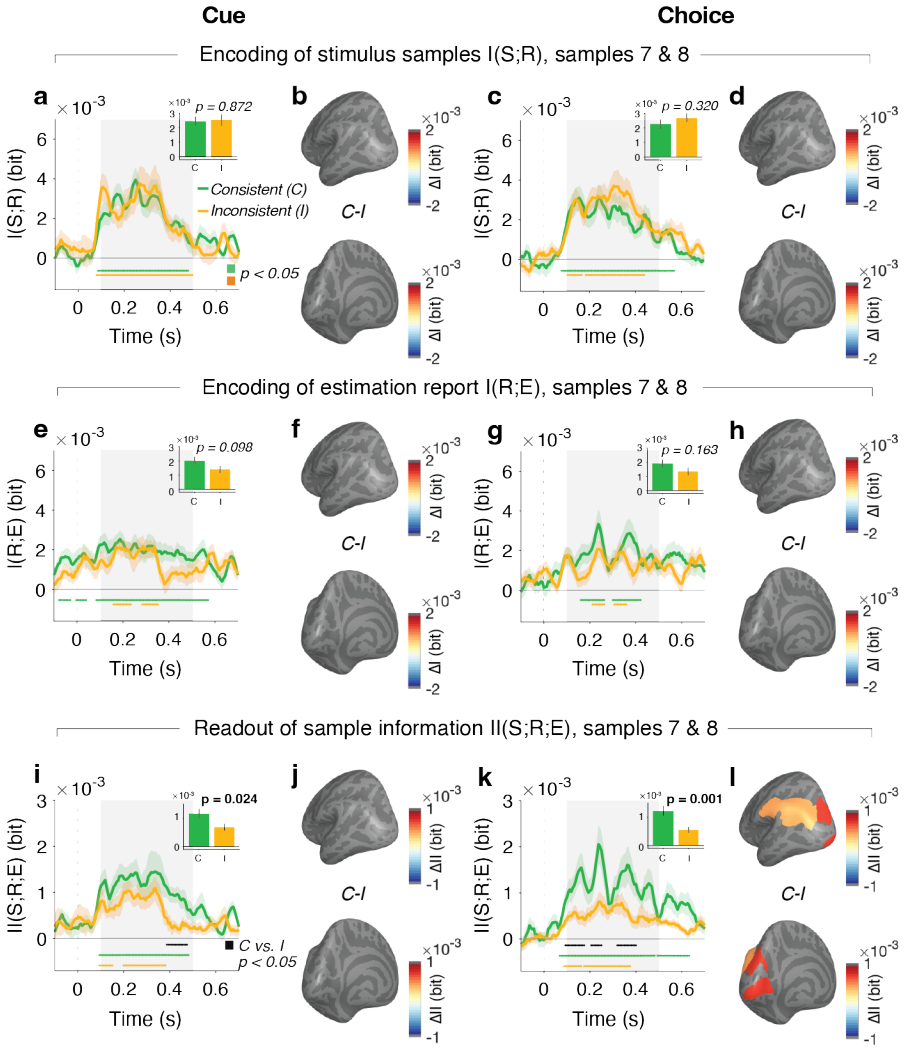

Figure 3. Effect of consistency on evidence encoding and evidence readout.

(a,c) Time courses of stimulus sample information I(S;R) in dorsal visual cortex for consistent (C) and inconsistent (I) samples, relative to the category of previous Cue (a) or Choice (c). Time-courses are pooled across samples 7 and 8, following the choice and exhibiting the behavioral consistency effect (see Figure 1e,f). Bar graphs, mean across gray-shaded interval with significant I(S;R) (compare Figure 2). **(b,d)** Corresponding statistical maps for the difference between information measures for consistent and inconsistent. **(e-h)** As (a-d), but for estimation information I(R;E). **(i-l)** As (a-d), but for intersection information II(S;R;E). Bars below time courses denote significant difference from pre-trial (-0.1 s – 0 s) baseline (colored), or significant difference between consistent and inconsistent. Time-courses were smoothed with a Gaussian kernel (sigma = 37.5 ms). Statistical tests for bar graphs and maps were done on non-smoothed data (mean across gray-shaded interval), using two-sided permutation tests; p-values cluster-corrected for time courses and FDR-corrected for maps (threshold: $p < 0.05$). N=29 participants for all panels; bars or lines, group mean; error bars or shaded areas, SEM.

In contrast, intersection information II(S;R;E) was larger for consistent than for inconsistent samples (Figure 3i-l). In line with the behavioral consistency effects in Figure 1f, g, the difference was prominent in the Choice condition (Figure 3k,l), and less robust (but significant for the 0.1 s - 0.5 s analysis time window in dorsal visual cortex, Figure 3i) in the Cue condition. Also in line with the behavioral effects, this neural effect was present for a short period following the choice (i.e., sample positions 7 and 8) with no effects for any other part of the trial (Figure S2 for corresponding samples from first trial interval and Figure S3 for non-thresholded maps across the complete trial). For Choice, the consistency effect on II(S;R;E) was present in a swath of parietal cortex including dorsal visual cortex and inferior parietal cortex as well as in primary visual cortex (V1; Figure 3l). No such significant effect of consistency survived multiple-comparison corrections in the Cue condition (Figure 3j and Figure S2 for non-thresholded maps).

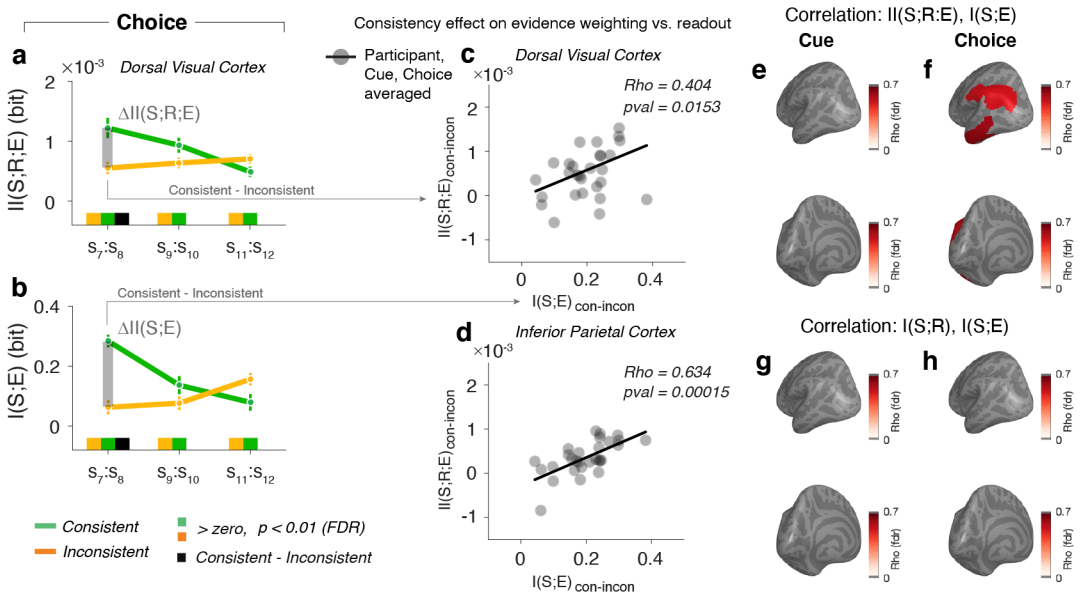

Figure 4. Consistency effect on intersection information in parietal cortex relates to individual behavioral bias.

(a) Behavioral consistency effect as difference in psychophysical kernels I(S;E) between consistent and inconsistent samples, pooled for two samples in second trial interval, shown for Choice condition for illustration. **(b)** Corresponding neural consistency effect as difference in intersection information II(S;R;E) between consistent and inconsistent samples in dorsal visual cortex (0.1 s - 0.5 s from sample-onset, corresponding to bar graphs from Figure 3k). Bars below time courses, statistical significance (permutation tests, FDR-corrected). **(c,d)** Across-participant correlations between behavioral and neural consistency effects for samples 7,8 (shaded in a,b) in dorsal visual cortex (c) and inferior parietal cortex (d). For scatter plots, consistency effects were averaged across Choice and Cue conditions. **(e,f)** Maps of across-subject correlations of behavioral versus neural consistency effects in (c,d), but separately for Cue (e) and Choice (f); one-sided test, threshold: $p < 0.05$, FDR-corrected. **(g,h)** As (e,f) but substituting stimulus information I(S;R) for II(S;R;E) as neural measure. N=29 participants for all panels; dots, group mean; error bars, SEM.

Our previous results are consistent with the notion that the readout of the sample information in parietal cortex (and V1) for the estimation report depended on the sample's consistency with the previous categorical choice or cue. If the effect of consistency on neural intersection information in Figure 3i-l contributed to the consistency-dependent behavioral weighting profile in Figure 1f,g, then the strengths of neural and behavioral consistency effects should be related. We, therefore, correlated each participant's difference in II(S;R;E), separately for all areas from Figure 3l, between consistent and inconsistent samples with the corresponding differences in I(S;E), illustrated as scatter plot for dorsal visual cortex in Figure 4c

(consistency effects pooled across Choice and Cue conditions). There was a significant correlation in the two parietal regions showing the consistency effect on intersection information, dorsal visual and inferior parietal cortex (Figure 4c,d). For the Choice (but not Cue) condition, this effect was robust with respect to multiple comparisons correction across the brain (Figure 4f). No such correlation was found for I(S;R) instead of II(S;R;E) (Figure 4g,h), highlighting the specificity of the brain-behavior correlation for intersection information.

Together with Figure 3, the results from Figure 4 support the idea that the bias of evidence processing induced by an intermittent choice is mediated by an altered readout of that evidence from parietal cortex: Evidence following the choice is read out from parietal cortex in a manner that depends on consistency; and the strength of this consistency effect in an individual reflects the individual's strength of consistency effect on behavioral evidence weighting.

DISCUSSION

Human judgment and decision-making are shaped by biases and heuristics[26–28]. One of the most prevalent of those is the tendency to process new evidence selectively, so as to confirm the beliefs and judgments one has previously committed to[7,29,30], whereby evidence supporting a previous decision is given more weight, and evidence contradicting that decision is given less weight in future judgments[8,28]. We reasoned that such selective processing of post-decisional evidence could come about either through a change in the encoding precision of evidence in sensory cortex, or a change in the readout of that sensory representation for subsequent decisions. Our information-theoretic dissection of evidence encoding and readout in patterns of human cortical population activity provide no evidence for the encoding scenario. Instead, our results clearly support the readout scenario, and implicate posterior parietal cortex as an important processing stage in the underlying neural pathway.

This result is important for unraveling the information processing mechanisms underlying choice-induced biases. Choice-induced biases can be advantageous under certain conditions[31] but may be detrimental in others. Maintaining sensory representations that are undistorted by categorical beliefs is beneficial for the observer. It can prevent the erroneous maintenance of beliefs states (hysteresis) when additional evidence or an actual change in source location warrants a switch in categorical belief[4]. Indeed, behavioral findings on retrospectively (i.e., after evidence presentation) induced choice biases show that perceptual estimation behavior after a forced change in choice due to feedback is best modeled by a switch in categorical belief, but unaltered sensory representations[32], in line with our current neural data. More generally, pinpointing the origin of confirmation bias to a modulation of evidence readout, rather than evidence encoding, suggests that confirmation biases may result from a deliberate decision process, and thus can potentially be counter-balanced by appropriate incentives.

Other work has identified post-decisional biases in other perceptual designs that also entail categorical and continuous estimation judgments, but, differently from our task, with both judgments prompted after termination of the evidence stream[8,18,33,34]. The biases in the continuous estimation that are here "retrospectively" induced by the categorical choice resemble established post-decisional biases in value-based decisions[35]. These results point to self-consistency as a general principle shaping human decision-making[18,36]. Here, we found that such choice-induced consistency effects also act "prospectively" and

pinpointed the readout of sensory evidence as the source of the biased processing. Whether or not this mechanism also underlies the retrospective post-decision bias described above remains to be tested in future work.

We used two complementary approaches to modulate participants' categorical beliefs about the source in the middle of the evidence stream: prompting them to report a judgment about that category or providing an external cue that informed about that category in a probabilistic manner. While both manipulations induced a detectable behavioral bias in subsequent evidence processing and in neural intersection information, that bias was more robust in the Choice context. Indeed, in the Choice context, but not the Cue context, there was a strong encoding of final estimation decision in frontal cortical regions during the second trial interval that followed the choice (Figure 2d,e). It is tempting to speculate that the act of making a categorical choice leaves a stronger "trace" in the brain that biases the use of subsequent information more prominently than an external (similarly informative) cue (Figure 1c), in line with the notion of self-confirmation in social psychology[7].

Our results suggest that the choice-induced bias characterized here is unlikely to be caused by a selective modulation of sensory representations[37–39] through feedback signals[40–42]. We found no evidence for a change of stimulus encoding in visual cortex, depending on the stimulus consistency with the previous choice or cue. Our manipulation of categorical beliefs with the external cue invites comparison with a large body of studies using cues to manipulate selective attention and measured neural activity in visual cortex. Many of those studies reported changes in the sensory representation via gain modulation[37]. This apparent discrepancy may be due to differences between our current approach and the previous studies of selective attention. For example, we presented imprecise cues that divided the continuous visual feature space into two coarse categories (left vs. right from reference line), whereas the cues have commonly been precise in most previous studies on attention mechanisms. Furthermore, we quantified sensory encoding in terms of the mutual information between stimulus and an MEG-derived estimate of the local field potential response, as opposed to single neuron firing rates in the previous studies. Finally, it is possible that the reported effects of attention on sensory encoding precision co-existed with stronger effects on sensory readout, which largely mediated the behavioral effects of attention. Indeed, model-based fMRI work points to exactly such a scenario for attentional selection[43]. It would be useful to apply our approach to a wider range of cueing tasks including those commonly used in the attention literature.

The interplay between encoding and readout of information, operationalized here as Intersection information, has been identified as an important criterion for identifying neural activity patterns guiding perceptual decisions[16] (see also[44–46]). The rationale is that not all information about the stimulus encoded in the brain necessarily informs behavior; likewise, not all neural activity shaping behavior is based on the readout of stimulus information (e.g., choice history biases[47] or motor biases). Because intersection information quantifies the information about a stimulus contained in a neural activity pattern that informs behavior, it enables the most specific statistical (i.e. correlative) delineation of putative task-relevant neural signals in the brain. Here, intersection information enabled us to isolate a clear effect of consistency that directly originates from the sample evidence that was neither evident in stimulus nor estimation information. Given the limits of recordable data, we averaged all information quantities across features (i.e., PCs) of the neural activity pattern within a given area (Figure 2a), rather than evaluating the joint

information across neural features. Thus, while unlikely, we cannot exclude the possibility that there was a consistency effect in the joint stimulus information across neural features.

A few previous studies have probed the neural basis of confirmation biases[29,30]. One reported behavioral (drift diffusion) model fits indicating that sensory evidence following an initial categorical judgment of a noisy stimulus was selectively accumulated depending on its consistency with the initial judgment, in a manner that depended on the observers' confidence about the initial judgment, a conclusion that was supported by the dynamics of decision-related (sensor-level) MEG signals[30]. Selective evidence accumulation may be due to a selective modulation of evidence encoding or a selective evidence readout[48]. In a social setting, fMRI responses in the posterior medial prefrontal cortex indicated failure to use other's opinion when they were disconfirmatory[29]. Neither of these studies have dissected the neural computations underlying selective evidence weighting by distinguishing between the encoding versus the readout of evidence for decisions, as we did here.

Our findings bear a striking analogy with studies of the mechanisms of visual perceptual learning in macaque monkey cortex[49,50]. This work uncovered an improvement of the readout of the sensory representation from visual motion-selective cortical area MT by downstream decision circuits during learning[49], an effect that could be modeled as a reinforcement-dependent optimization of readout weights[50]. Our current work establishes a similar change in sensory readout weights, but on a substantially more rapid timescale, at the level of trial-to-trial and even sample-to-sample (within trial) resolution. This analogy points to the possibility of systematically changing individuals' tendencies toward self-confirmation through directed learning strategies.

In conclusion, adopting categorical beliefs about the state of the environment does not affect the encoding of subsequent evidence in the human brain, but selectively biases the readout of the encoded information for future decisions. Because confirmation bias is prevalent in cases of real-world significance, our insights have important practical implications. For example, evidence readout is more susceptible to strategic interventions and learning of the decision-maker than evidence encoding. Our results indicate that the room for manipulating self-confirmation tendencies in real-life decisions may be substantial.

METHODS

Participants

34 healthy human volunteers (sex: 16 male, 18 female) aged between 19-39 years (mean = 27 years, SD = 5) participated in the study. All participants gave written consent and were naive to the objectives of the study. Participants self-reported normal or corrected-to-normal vision and no history of psychiatric or neurological diagnosis. They received remuneration in the form of an hourly rate of 10 €, and a bonus (25 €) for completing all planned sessions. The study was approved by the ethics review board of the Hamburg Medical Association responsible for the University Medical Center Hamburg-Eppendorf (UKE).

Four participants were excluded during the MEG preprocessing or source reconstruction stage (i.e., prior to the tests of the main experimental effects, see below), leaving N=30 participants. Three of those showed persistent MEG artifacts (of unclear origin) at the sensor-level, which could not be removed with the preprocessing pipeline described below; in a fourth, the source reconstruction (see below) failed for several parcels. One additional participant did not have a sufficient trial count for quantifying neural

intersection information separately for consistent and inconsistent conditions (see below, *Information-theoretic analyses of behavior and neural activity.*), leaving N=29 for Figures 3 and 4. We ensured that the behavioral consistency effects (i.e., on I(S;E)) shown in Figure 1 (panels f-g) were unaffected when restricting the analysis to the same N=29 participants included in Figures 3 and 4.

Stimuli

Stimuli were generated using Psychtoolbox-3 for MATLAB and were back-projected on a transparent screen using a Sanyo PLC-XP51 projector at 60 Hz during MEG recordings, or on a VIEWPixx monitor during the training session in a behavioral psychophysics laboratory. The task-relevant stimulus, so-called evidence samples, were small, checkered discs (0.8° diameter, temporal frequency, 6.7 Hz; spatial frequency, 2°) whose eccentricity was fixed at 5° of visual angle and whose polar angle varied from sample to sample within each trial. Each sample was presented for 100 ms followed by 50 ms blank. Two placeholders were present throughout each trial: a fixation cross at the center of the screen with varying orientation and color informing participants about trial intervals, and a light-gray reference line extending downward (or upward) from fixation to 5° eccentricity (Figure 1a). In different trial blocks, the sample positions were restricted to the upper or lower hemifield (counterbalanced). The orientation of the checkerboard pattern changed according to the sample position. Stimuli were presented against a gray background.

Definition of sensory evidence

We defined sensory evidence as sample positions with respect to the reference line: for both upper hemifield blocks and lower hemifield blocks, a sample position of 0° indicated that discs were positioned on the vertical meridian (i.e., aligned with the reference line), positive angles indicated locations right from the reference line, and negative angles indicated locations left from the reference line. In each trial, the sample positions were drawn from a Gaussian distribution, with a (generative) mean that varied from trial to trial and fixed standard deviation of 20°. Generative means for each trial were sampled from a uniform distribution (-14°, +14° relative to reference line in 1° increments -15°, +15°, in 5° increments for the first 3 participants). All samples were constrained to the range of -40° to + 40° from the generative mean.

Behavioral tasks

Participants were asked to monitor a total of 12 evidence samples, infer their hidden source (i.e., generative mean), and report that as a continuous estimate at the end of the trial. After the first six samples, participants were also asked to report a categorical judgment of the evidence source (left or right from reference line; Choice condition); or to make a stereotypical button press that was pre-assigned for each block of trials (Cue condition). In the latter task, the button press was followed by an imprecise visual cue indicating the most likely (75% validity) category of the evidence source. This cue was as an arc spanning 14° of polar angle (i.e., the range of generative means on either side of the reference line). Participants were instructed to maintain fixation throughout the trial.

Each trial started with a baseline interval ranging between 0.5 s – 1 s. The intermittent response time window was fixed at 1.25 s, the cue display period was fixed at 0.5 s immediately after the response window. After a short delay period of 0.1 s – 0.3 s, the fixation cross rotated to an 'x' to indicate the estimation period. The estimation response window ended with the participants' estimation report or at 5 s, whichever comes first. The inter-trial-intervals were drawn from a uniform distribution (3.5 s, 5 s). Cumulative feedback about the binary choice (percentage correct; if the generative mean was zero, the correctness was assigned based on randomly choosing left/right as correct), and the estimation (root mean squared deviation of estimation from the generative mean) were provided at the end of each block. In addition, feedback about the number of trials they have missed the button press or estimation, as well as the number of trials in which they have failed to maintain fixation was also provided.

Experimental procedure

All participants completed five sessions on separate days: one behavioral training session in a psychophysics laboratory where participants were instructed about, and familiarized with, the task; and four main experimental sessions in an MEG laboratory.

Behavioral training session. On the first visit, participants were briefed about the purpose of the study and measurement. Since the measurements took place during the COVID-19 pandemic, participants submitted negative tests and were informed about additional hygiene measures implemented in order to minimize risk from the virus. Experimenters also were required to test negative before any measurements during this period. For the training, participants performed a behavior-only version of the main task broken down into seven steps, with their head fixed on a chinrest, 60 cm from the monitor. The training started with practicing making a left/right choice by accumulating the 12 checkered dot positions presented sequentially. The final step of the training was similar to the main task except that the intermittent button press was preassigned and there was no cue. The training lasted around 2 hours. Participants were trained both with and without feedback about their performances on a trial-by-trial basis. Participants' performance was checked immediately after data collections.

MEG sessions. Participants performed two sessions of the Cue task, followed by two sessions of the Choice task, in that order, on separate days. Each session consisted of 8 blocks (58 trials each). In addition, visual cortex localizer blocks were added at the end of each session for both upper and lower hemifield stimuli. Participants were seated 61 cm from the screen during MEG recording. Before each session, participants underwent a brief practice block to familiarize themselves with the joystick control in the MEG environment. Each session contained around 2 h of recording time, resulting in a total of approximately 6-hour measurement per task. Identical trials were used for both tasks and all participants. In addition, there were 4 identical trials for each session, 3 in the exact sequential order, and one shuffled across the 12 sample positions. We collected 1856 trials (464 trials x 2 sessions per task) in total from each participant during MEG.

Data acquisition

MEG data were acquired with 275 axial gradiometers (equipment and acquisition software, CTF Systems) in a magnetically shielded room (sampling rate, 1200 Hz). Participants were asked to minimize any

movements. Head location was recorded in real-time using fiducial coils at the nasal bridge and each ear canal. A template head position was registered at the beginning of the first session, and the participant was guided back into that position before each subsequent task block. Ag–AgCl electrodes measured the ECG and EOG which was monitored throughout the session. Eye movements and pupil diameter were recorded at 1000 Hz with an EyeLink 1000 Long Range Mount system (equipment and software, SR Research). On a separate day, T1-weighted MRIs were acquired to generate individual head models for source reconstruction. Behavioral data (button presses and joystick displacement for estimation reports) were collected simultaneously with the electrophysiological data.

Data analysis

Data were analyzed in MATLAB (MathWorks) and Python using a combination of custom scripts and the following toolboxes: FieldTrip, MNE python, BROJA_2PID toolbox[51] and in-house pipeline for source-reconstruction PyMEG and custom-made scripts. pymeg (https://github.com/DonnerLab/pymeg).

Trial exclusion. Some trials were excluded from analysis based on criteria related to behavior and MEG data. Trials with missed intermittent button press (choice condition) or missed final estimation were excluded. Also, trials with final estimation reaction time shorter than 300 ms were excluded, since these trials were deemed too fast to have made a decision, considering the fact that the estimation cursor was always located aligned to the vertical reference line. More specifically, trials that could induce a zero generative mean bias were excluded. Based on MEG data, trials with the following artifacts were excluded: sensor noise, muscle artifacts, eye-movement related artifacts (blinks, saccades). Trials used in both behavioral and MEG analyses were matched. On average, 22.1% of trials (standard deviation: ±10.7%) were excluded per individual dataset, leaving 723±99 (mean and standard deviation) trials per subject for the analyses described below.

Analysis of behavioral data. Estimation reports were plotted as a function of the generative or the sample means for each trial. To this end, the sample means were binned into 21 bins with 2° increments. Performance of the intermittent categorical judgment (Choice task only) was evaluated by fitting a psychometric function logistic function as a function of the sample mean for the first 6 samples.

MEG preprocessing. Continuous data were segmented into task blocks, high-pass filtered (zero phase, forward-pass FIR) at 0.1 Hz and band-stop filtered (two-pass Butterworth) around 50 Hz, 100 Hz and 150 Hz to remove line noise. The data was down-sampled to 400 Hz. Trials containing the following artifacts were discarded: translation of any fiducial coil > 6 mm from the first trial, blinks (detected by the EyeLink algorithm), saccades > 1.5° amplitude (velocity threshold, 30° s^{-1}; acceleration threshold, 2,000° s^{-2}), squid jumps (detected by Grubb's outlier test for intercepts of lines fitted to single-trial log-power spectra), sensors with data range > 7.5 pT and muscle signals (after applying a 110 Hz – 140 Hz Butterworth filter and z scoring) of z > 10. In addition, trials containing artifacts due to cars passing by in the vicinity of the building were detected and discarded. The remaining data for that recording were then subjected to temporal independent component analysis (infomax algorithm), and components containing blink or heartbeat artifacts were identified manually and removed.

Source reconstruction. We used linearly constrained minimum variance (LCMV) beamforming to estimate activity time courses at the level of cortical sources. For source reconstruction, data were

segmented into epochs ranging from 0.5 s before 1st sample onset and 4.5 s after, which includes the two evidence streams and the intermittent button press (and cue, for Cue condition). We constructed individual three-layer head models (except for three participants, for whom we used one-layer models) from structural MRI scans using FieldTrip. Head models were aligned to the MEG data by a session-specific transformation matrix generated with MNE. Transformation matrices were generated using MNE software. We reconstructed individual cortical surfaces from MRIs using FreeSurfer and applied the parcellation from an aligned anatomical atlas[20] to each surface. We used MNE to compute LCMV filters confined to the cortical surface (4,096 vertices per hemisphere, recursively subdivided octahedron) from the covariance matrix computed across all time points of the cleaned and segmented data. The covariance matrix was averaged across all trials per session. For each vertex location, we determined the source orientation yielding maximum power by means of singular value decomposition. The LCMV filters were used to project single-trial broadband MEG time series into source space. We computed MEG responses at each vertex by first aligning the polarity of time series at neighboring vertices, since the beamformer output potentially included arbitrary sign flips for different vertices.

Parcellation and regions of interest (ROIs). We used the atlas from Glasser and colleagues[20] as parcellation of the cortical surface into 180 well-defined areas based on multimodal (functional and anatomical) data. All MEG data analyses described below were first performed separately for each of these 180 areas (see Figure 2a). Because the dimensionality of source-level MEG data is typically smaller than 180, we next averaged the resulting information theoretic measures across all areas from each of 22 area groups described in reference[20]. Complementary to comprehensive mapping of all effects across all the complete set of area groups, our analyses focused on two of those groups as specific ROIs. We chose "Dorsal Stream Visual Cortex" (referred to as "dorsal visual cortex" for brevity)[20] as *a priori* ROI due to its established role in visual-spatial behavior[23–25] for Figures 2-4. Dorsal visual cortex is a group of visual field maps within and around the intraparietal sulcus, comprising the following areas: V3AB, V6, V6A, V7, and IPS)[21,22]. To depict correlations between individual effects of consistency on behavioral evidence weighting versus neural intersection information in Figure 4, we also chose the "Inferior Parietal Cortex" group (comprising areas PGp, IP0-2, PF, PFt, PFop, PFm, PGi, PGs) as an additional ROI. This selection was based on the significant consistency effect on intersection information in that group.

Dimensionality reduction. We used principal component analysis (PCA) across all vertices and trials, within each parcel to reduce the dimensionality of the data within each of the 180 parcels The source data was down-sampled to 200 Hz, which resulted in a 1001 sample segment for a trial (-0.5 s ~ 4.5 s; 0 s: 1st sample onset). We then chose *M* PCs which cumulatively explained 90 % of the variance in the data (Figure 2a). The number of selected PCs for each parcel ranged from 1 to 15. All analyses reported in the following were done on these PCs.

Information-theoretic analyses of behavior and neural activity. We used mutual information to quantify the amount of information encoded in neural activity (R) about stimulus samples S or subjects' behavioral estimation reports E, as a measure of sensory evidence (or report) encoding in neural activity. Likewise, we used mutual information to quantify the relationship between behavioral estimation reports E and the stimulus samples S. Here, S was the sensory evidence sample expressed (angular position with respect to reference line; see *Definition of sensory evidence* above), and E was the participant's estimation report expressed in the same reference frame. Mutual information quantifies the effect of all the single-trial

(both linear and nonlinear) statistical relationship between two random variables[52,53]. For each parcel (cortical area) and task-relevant variable X (i.e., S or E), we computed mutual information carried by $R_{i,t}$, that is, the activity of each PCi (see above, *Dimensionality reduction*) and each time point t, as follows:

$$I(X; R_{i,t}) = \sum_{x, r_{i,t}} p(x, r_{i,t}) \, log_2 \frac{p(x, r_{i,t})}{p(x) p(r_{i,t})}, \quad (1)$$

where $p(x, r_{i,t})$ was the joint probability of observing a neural or behavioral response $r_{i,t}$ and for a value x of the task-relevant variable, and $p(r_{i,t})$ and $p(x)$ were the marginal probabilities of $r_{i,t}$ and x respectively. The sum is over all possible values of $r_{i,t}$ and x. We discretized the values of neural responses $R_{i,t}$ (independently for each time point and PC) and behavioral variables X (both S and E were continuous) across trials into M equally populated bins[54]. Equally populated binning is advantageous for calculations of information about continuous-valued variables because it maximizes the marginal entropy of the variables given the number of bins and this often increases the mutual information values[54]. We then estimated $p(x, r_{i,t})$ in Eq (1) from data using the direct method[55,56] (i.e., counting the occurrences of each possible pair of $(r_{i,t}, x)$ values across trials and dividing by the total number of trials).

We chose M=3 bins for all main analyses reported in this paper because this yielded a reasonable number of observations per bin (and thus more accurate and unbiased information estimates[55,57]) for all analyses, including the tri-variate probability distributions required for computing intersection information (Figure S4a,e)[55,56,58], even after splitting data by consistency. For the inconsistent stimulus samples, the condition with the smallest data yield, M=3 bins yielded on average five observations per bin in the Cue condition and four observations per bin in the Choice condition, respectively. Importantly, the qualitative patterns (time courses and cortical distributions) of all information theoretic measures shown in Figure 2 were stable over a range of bin numbers (M=3, 5, 7, 9) for both task conditions (Figure S4), supporting our choice of M=3.

To correct for the bias in information estimation due to limited sampling, we subtracted the average information after shuffling X across experimental trials from information-theoretic measures (N=100 shuffles for mutual information, N=20 shuffles for intersection information) shown in Figure 2, as this correction removes the bias conservatively[55,57]. Further, to eliminate confounding effects of possible residual errors in information estimation (which depend on the number of available trials[55]) when comparing mutual information values across conditions, we matched the number of trials by random subsampling (N=5) of the condition with more trials when computing information.

We used intersection information (II)[17] to quantify the amount of stimulus information encoded about the stimulus S by the neural signal $R_{i,t}$ that informed the behavioral estimation report E. This measure is computed from the tri-variate probabilities p(s, $r_{i,t}$, e) of observing stimulus s, estimation c, and neural response $r_{i,t}$ at time t in location i in the same trial. II quantifies the part of information in neural responses that is common to both stimulus and estimation information. II is computed as the minimum between two terms with similar but slightly different interpretations, as follows:

$$II(S; R_{i,t}; E) = min[SI(E; \{S, R_{i,t}), SI(S; \{E, R_{i,t})], \quad (2)$$

The first term is $SI(E;\{S,R_{i,t})$, the information about estimation E shared between stimulus S and neural response $R_{i,t}$. This quantity is obtained (using the definition of shared information derived in[59]) by computing the maximum shared dependency between S and R conditioned on E, defined as the mutual information between stimuli and neural responses minus the mutual information between stimuli and responses conditioned on choice, with the maximum computed over the space of all distributions q(s, $r_{i,t}$,e) that preserve the marginals p(s,e) and p($r_{i,t}$,e). The second term is $SI(S;\{E,R_{i,t})$, the information about S shared between stimulus E and $R_{i,t}$. This term is obtained in a similar way by computing the maximum shared dependency between C and $R_{i,t}$, conditioned on S. Minimizing between the two terms ensures that II satisfies key properties that would be expected from a measure with this interpretation, including that independent S and $R_{i,t}$, that II is non-negative and bounded by the stimulus and estimation report information encoded in neural activity, and by the information between stimulus and estimation report. We again discretized $S, R_{i,t}$ and E into 3 bins (see below), we matched the number of trials when comparing II between conditions.

Psychophysical kernels quantify the impact of stochastic fluctuations in sensory evidence samples on behavioral reports[19]. Our design entailed systematic trial-to-trial variations in the generative mean, which were important for demonstrating stimulus-guided choice and estimation behavior (Figure 1d,e) and keeping participants engaged, but introduced correlations stimulus samples, categorical cue or choice, and estimation reports. In the ideal, the analysis of psychophysical kernels would focus only on trials with generative means of 0°. In practice, it was necessary to use a small range of generative means around 0°, so as to obtain a sufficiently large number of trials for the key analyses of consistency effects (small counts of inconsistent samples). The range of means from -4° to +4° was the smallest range that yielded sufficient numbers of observations necessary for the computation of II(S;R;E). We verified through simulations of an synthetic observer that this range did not yield spurious consistency effects in I(S;E). The agent was presented with the exact same sequences of evidence seen by our participants, and made intermittent choices based on the sign of the mean of samples 1-6, and continuous estimations based on the mean of all 12 samples (i.e., "perfect integrator"). The agent was simulated for (i) varying levels of "decision noise" (fractions of trials with random judgements) giving rise to varying levels of choice accuracy and (ii) varying subsets of trials with different ranges of generative means surrounding 0°. As for the actual participants (Figure 1f,g), we then computed I(S;E) after sorting samples as consistent or inconsistent with respect to the intermittent choice. The stimulations showed no difference between consistent and inconsistent conditions (which would have been indicative of a spurious effect because the simulated agent was unbiased) the range of means from -4° to +4°, for choice accuracies (i.e., noise levels) matching participants' data (Figure 1c). Furthermore, when spurious consistency effects occurred for larger ranges of generative means and/or larger choice accuracies, those were present both in the second and in the first interval, unlike what we observed in the data (Figure 1f,g). We used the same subset of trials for all analyses of the effects of consistency on neural information measures (Figures 3,4). We restricted the computations of information measures to participants with at least 3 observations per bin (i.e., 81 trials for II(S;R;E) see above) for the subset of trials with generative means between -4° to 4°.

Statistical analysis. For all statistical tests, we used a permutation-based procedure (number of permutations: 2000, otherwise specified) in order to test effects between experimental conditions. The thresholds corresponded to p = 0.01 or higher when testing against baseline or p = 0.05 when testing

between two conditions. For time courses, we used a cluster-based permutation test (clusters in the time domain). For statistical maps, we averaged information measures across the interval 0.1 s - 0.5 s from sample onset and performed permutation tests for each area group (average across hemispheres and parcels per area, see above), and used FDR correction to control for multiple comparisons.

Data availability

Raw behavioral data and preprocessed MEG data will be made available upon acceptance.

Code availability

All analysis code to reproduce the reported results and figures from the shared data will be made available on GitHub (https://github.com/DonnerLab) upon acceptance.

REFERENCES

1. Behrens, T. E. J., Woolrich, M. W., Walton, M. E. & Rushworth, M. F. S. Learning the value of information in an uncertain world. *Nat Neurosci* **10**, 1214–1221 (2007).
2. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron* **92**, 1398–1411 (2016).
3. Gold, J. I. & Stocker, A. A. Visual Decision-Making in an Uncertain and Dynamic World. *Annu. Rev. Vis. Sci.* **3**, 227–250 (2017).
4. Murphy, P. R., Wilming, N., Hernandez-Bocanegra, D. C., Prat-Ortega, G. & Donner, T. H. Adaptive circuit dynamics across human cortex during evidence accumulation in changing environments. *Nat Neurosci* **24**, 987–997 (2021).
5. Tavoni, G., Doi, T., Pizzica, C., Balasubramanian, V. & Gold, J. I. Human inference reflects a normative balance of complexity and accuracy. *Nat Hum Behav* **6**, 1153–1168 (2022).
6. Bacon, Francis. Novum organum. in *The English philosophers from Bacon to Mill* 24–123 (New York: Random House, 1939).
7. Nickerson, R. S. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* **2**, 175–220 (1998).
8. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M. & Donner, T. H. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Current Biology* **28**, 3128-3135.e8 (2018).
9. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review* **113**, 700–765 (2006).
10. Glaze, C. M., Kable, J. W. & Gold, J. I. Normative evidence accumulation in unpredictable environments. *eLife* **4**, e08825 (2015).
11. Brody, C. D. & Hanks, T. D. Neural underpinnings of the evidence accumulator. *Current Opinion in Neurobiology* **37**, 149–157 (2016).

12. Wilming, N., Murphy, P. R., Meyniel, F. & Donner, T. H. Large-scale dynamics of perceptual decision information across human cortex. *Nat Commun* **11**, 5109 (2020).

13. Wang, X.-J. Theory of the Multiregional Neocortex: Large-Scale Neural Dynamics and Distributed Cognition. *Annu. Rev. Neurosci.* **45**, 533–560 (2022).

14. Steinemann, N. A. *et al. Direct Observation of the Neural Computations Underlying a Single Decision*. http://biorxiv.org/lookup/doi/10.1101/2022.05.02.490321 (2022) doi:10.1101/2022.05.02.490321.

15. Baillet, S. Magnetoencephalography for brain electrophysiology and imaging. *Nat Neurosci* **20**, 327–339 (2017).

16. Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E. & Fellin, T. Cracking the Neural Code for Sensory Perception by Combining Statistics, Intervention, and Behavior. *Neuron* **93**, 491–507 (2017).

17. Pica, G. *et al.* Quantifying how much sensory information in a neural code is relevant for behavior. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 3686–3696 (Curran Associates, Inc., 2017).

18. Luu, L. & Stocker, A. A. Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife* **7**, e33334 (2018).

19. Waskom, M. L., Okazawa, G. & Kiani, R. Designing and Interpreting Psychophysical Investigations of Cognition. *Neuron* **104**, 100–112 (2019).

20. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

21. Wandell, B. A., Dumoulin, S. O. & Brewer, A. A. Visual Field Maps in Human Cortex. *Neuron* **56**, 366–383 (2007).

22. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic Maps of Visual Topography in Human Cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).

23. Silver, M. A., Ress, D. & Heeger, D. J. Topographic Maps of Visual Spatial Attention in Human Parietal Cortex. *Journal of Neurophysiology* **94**, 1358–1371 (2005).

24. Silver, M. A. & Kastner, S. Topographic maps in human frontal and parietal cortex. *Trends in Cognitive Sciences* **13**, 488–495 (2009).

25. Schluppeck, D., Curtis, C. E., Glimcher, P. W. & Heeger, D. J. Sustained Activity in Topographic Areas of Human Posterior Parietal Cortex during Memory-Guided Saccades. *J. Neurosci.* **26**, 5098–5108 (2006).

26. Tversky, A. & Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science* **185**, 1124–1131 (1974).

27. Kahneman, D. Maps of Bounded Rationality: Psychology for Behavioral Economics. *The American Economic Review* **93**, 1449–1475 (2003).

28. Gigerenzer, G. & Gaissmaier, W. Heuristic Decision Making. *Annu. Rev. Psychol.* **62**, 451–482 (2011).

29. Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R. & Sharot, T. Confirmation bias in the utilization of others' opinion strength. *Nat Neurosci* **23**, 130–137 (2020).

30. Rollwage, M. *et al.* Confidence drives a neural confirmation bias. *Nat Commun* **11**, 2634 (2020).

31. Qiu, C., Luu, L. & Stocker, A. A. Benefits of commitment in hierarchical inference. *Psychological Review* **127**, 622–639 (2020).

32. Luu, L. & Stocker, A. A. Categorical judgments do not modify sensory representations in working memory. *PLoS Comput Biol* **17**, e1008968 (2021).

33. Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).

34. Rollwage, M. & Fleming, S. M. Confirmation bias is adaptive when coupled with efficient metacognition. *Phil. Trans. R. Soc. B* **376**, 20200131 (2021).

35. Brehm, J. W. Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology* **52**, 384–389 (1956).

36. Stocker, A. A. & Simoncelli, E. A Bayesian Model of Conditioned Perception. in *Advances in Neural Information Processing Systems* (eds. Platt, J., Koller, D., Singer, Y. & Roweis, S.) vol. 20 (Curran Associates, Inc., 2008).

37. Maunsell, J. H. R. & Treue, S. Feature-based attention in visual cortex. *Trends in Neurosciences* **29**, 317–322 (2006).

38. Reynolds, J. H. & Heeger, D. J. The Normalization Model of Attention. *Neuron* **61**, 168–185 (2009).

39. Prat-Ortega, G. & Rocha, J. D. L. Selective Attention: A Plausible Mechanism Underlying Confirmation Bias. *Current Biology* **28**, R1151–R1154 (2018).

40. Nienborg, H. & Cumming, B. G. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature* **459**, 89–92 (2009).

41. Lange, R. D., Chattoraj, A., Beck, J. M., Yates, J. L. & Haefner, R. M. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *PLoS Comput Biol* **17**, e1009517 (2021).

42. Desimone, R. & Duncan, J. Neural Mechanisms of Selective Visual Attention. *Annu. Rev. Neurosci.* **18**, 193–222 (1995).

43. Pestilli, F., Carrasco, M., Heeger, D. J. & Gardner, J. L. Attentional Enhancement via Selection and Pooling of Early Sensory Responses in Human Visual Cortex. *Neuron* **72**, 832–846 (2011).

44. Yates, J. L., Park, I. M., Katz, L. N., Pillow, J. W. & Huk, A. C. Functional dissection of signal and noise in MT and LIP during decision-making. *Nat Neurosci* **20**, 1285–1292 (2017).

45. Zhao, Y., Yates, J. L., Levi, A. J., Huk, A. C. & Park, I. M. Stimulus-choice (mis)alignment in primate area MT. *PLoS Comput Biol* **16**, e1007614 (2020).

46. Valente, M. *et al.* Correlations enhance the behavioral readout of neural population activity in association cortex. *Nat Neurosci* (2021) doi:10.1038/s41593-021-00845-1.

47. Urai, A. E. & Donner, T. H. Persistent activity in human parietal cortex mediates perceptual choice repetition bias. *Nat Commun* **13**, 6015 (2022).

48. Urai, A. E., de Gee, J. W., Tsetsos, K. & Donner, T. H. Choice history biases subsequent evidence accumulation. *eLife* **8**, e46331 (2019).

49. Law, C.-T. & Gold, J. I. Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nat Neurosci* **11**, 505–513 (2008).

50. Law, C.-T. & Gold, J. I. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nat Neurosci* **12**, 655–663 (2009).

51. Makkeh, A., Theis, D. & Vicente, R. BROJA-2PID: A Robust Estimator for Bivariate Partial Information Decomposition. *Entropy* **20**, 271 (2018).

52. Quian Quiroga, R. & Panzeri, S. Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci* **10**, 173–185 (2009).

53. Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technical Journal* **27**, 379–423 (1948).

54. Magri, C., Whittingstall, K., Singh, V., Logothetis, N. K. & Panzeri, S. A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC Neurosci* **10**, 81 (2009).

55. Panzeri, S., Senatore, R., Montemurro, M. A. & Petersen, R. S. Correcting for the Sampling Bias Problem in Spike Train Information Measures. *Journal of Neurophysiology* **98**, 1064–1072 (2007).

56. Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R. & Bialek, W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.* **80**, 197–200 (1998).

57. Panzeri, S. & Treves, A. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems* **7**, 87–107 (1996).

58. Ince, R. A. A., Mazzoni, A., Bartels, A., Logothetis, N. K. & Panzeri, S. A novel test to determine the significance of neural selectivity to single and multiple potentially correlated stimulus features. *Journal of Neuroscience Methods* **210**, 49–65 (2012).

59. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. & Ay, N. Quantifying Unique Information. *Entropy* **16**, 2161–2183 (2014).

Acknowledgments

We thank Maxime Maheu, Joshua Calder-Travis, Ruud van den Brink, Anke Braun, Gina Monov, Marlene Petersson, and Emma Krink for help with data collection.

Author contributions

Conceptualization: BCT, AAS, THD; methodology, software, investigation: HP, AA, BCT, MC, SP, AAS, THD; data curation: HP, AA, BCT; formal analysis, visualization: HP; writing (first draft): HP, THD; writing (review and editing): HP, AA, BCT, MC, SP, AAS, THD; supervision: THD; funding acquisition: SP, AAS, THD; project management: THD.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project DO1240_2-2 (to THD), the German Federal Ministry of Education and Research (BMBF) project 01GQ1907 (to THD), the National Science Foundation (NSF), project IIS-1912232 (to AAS), and by the NIH Brain Initiative, grant R01NS108410 (SP).

Supplementary Figures
Confirmation Bias through Selective Use of Evidence in Human Cortex

Hame Park, Ayelet Arazi, Bharath Chandra Talluri, Marco Celotto, Stefano Panzeri,
Alan A Stocker & Tobias H Donner

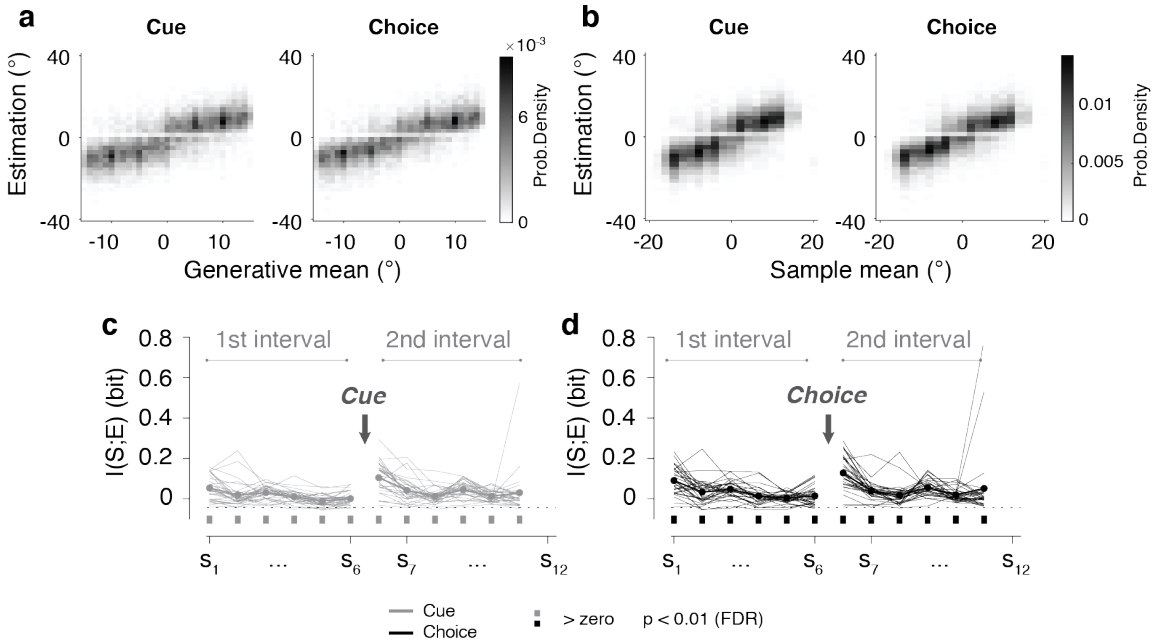

Figure S1. Quantification of estimation behavior.

(a, b) Distributions of estimation reports as a function of generative mean (a) and sample mean (b). Estimation reports were binned from -40° to 40°, in 2° increments. Individual probability densities were pooled across participants (N = 30). **(c, d)** I(S;E), the mutual information between sample (S) and estimation report (E), for Cue (c) and Choice (d) conditions. I(S;E), quantifies the impact of stochastic trial-to-trial fluctuations of sensory evidence at a given position in the sample sequence on behavioral estimation reports ("psychophysical kernels"[1–3]). Bars below time courses, statistical significance (permutation tests, FDR-corrected, threshold $p < 0.01$). In both conditions, I(S;E) was significantly larger than 0 for all sample positions, consistent with accumulation of evidence across the sequence and with the psychophysical kernels obtained in categorical decisions[1–3]. Note that I(S;E) values computed for all trials are smaller than the averages of I(S;E) values computed separately for consistent and inconsistent samples (Figure 1f,g). This is expected because the information contained in a pooled set of observations with inhomogeneous encoding properties (here: consistent and inconsistent samples pooled) is smaller than the average of the information contained in each set[4]. N=30 participants for all panels.

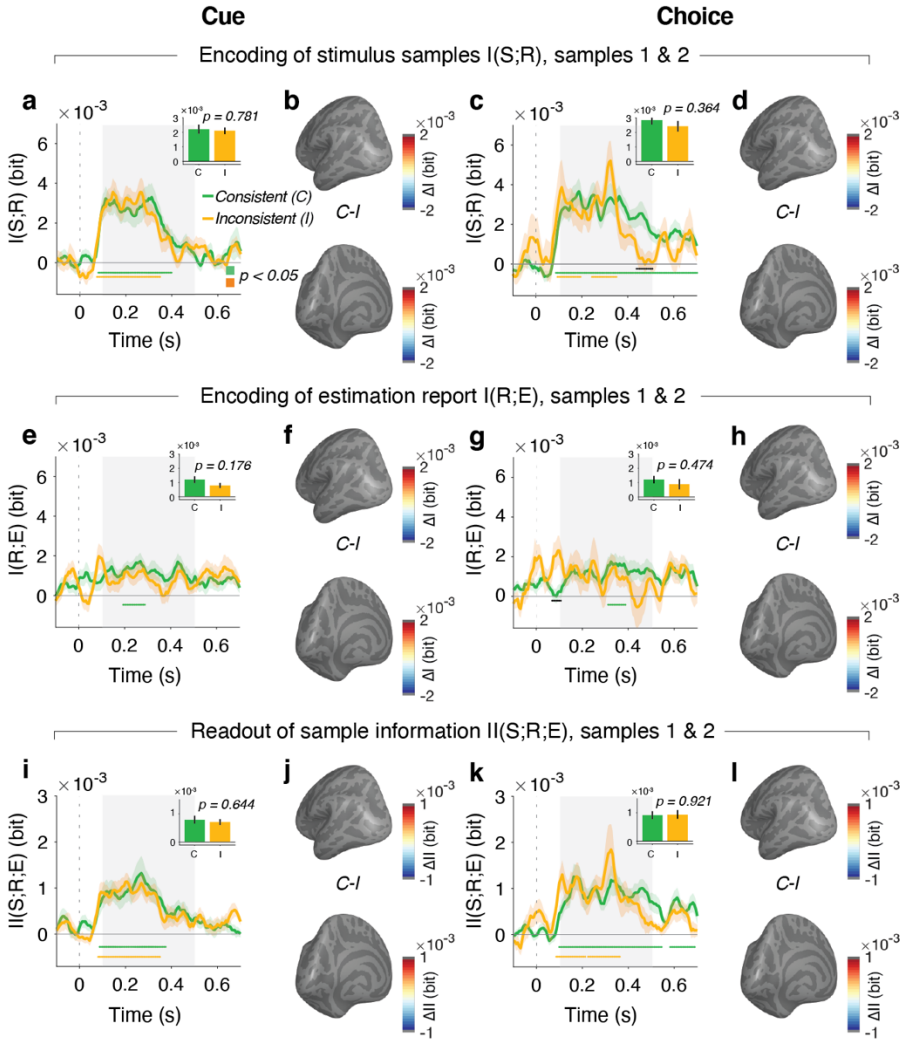

Figure S2. Effects of consistency with choice or cue on information measures during the first samples.

As main Figure 3, but for the first two samples of the first trial interval preceding choice or cue. Despite robust stimulus information in both conditions (Cue, Choice) for consistent (C) as well as inconsistent (I) samples, there is no effect of consistency on these measures anywhere, demonstrating the temporal specificity of the effect shown in Figure 3. N ranges between 27 and 29 participants for different panels, contingent on participant inclusion based on the criterion of at least 81 trials (for computation of intersection information) per condition and sample position (Methods).

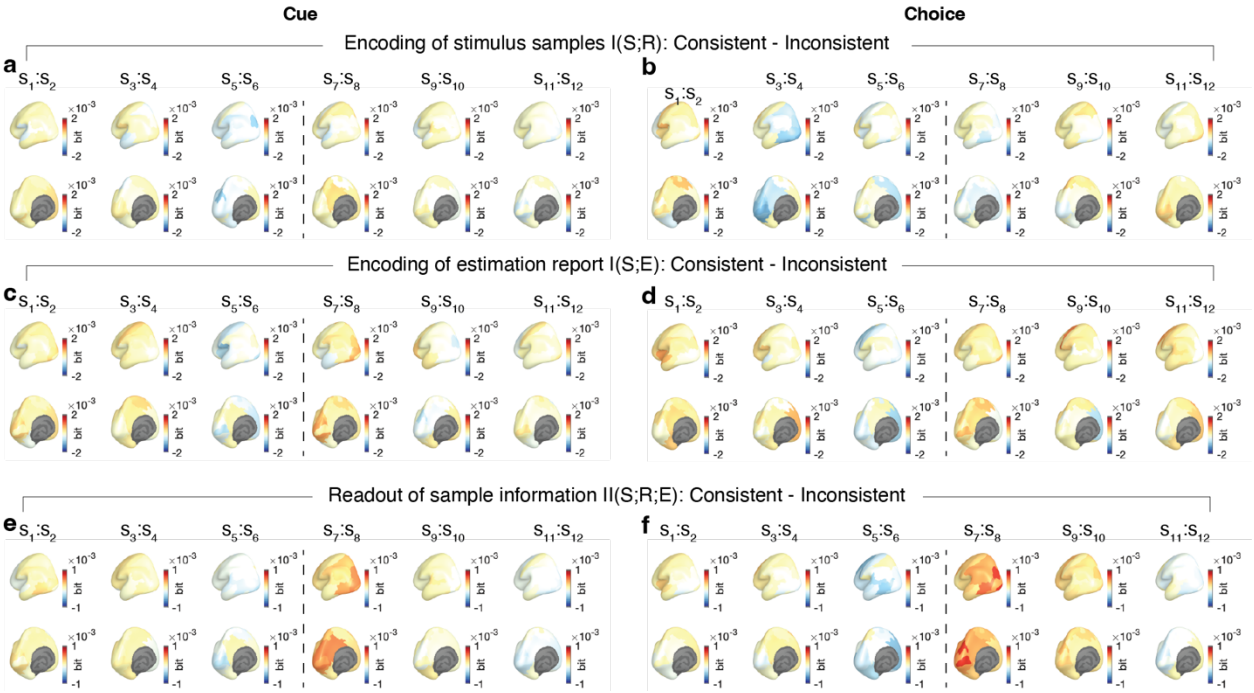

Figure S3. Effects of consistency with choice/cue on information measures across the trial.

Non-thresholded maps of consistency effects (consistent - inconsistent) on information measures across the complete trial (two successive samples pooled, see main text), for Cue (left) and Choice (right) conditions, respectively. **(a,b)** Effects on I(S;R). **(c,d)** Effects on I(R;E). **(e,f)** Effects on II(S;R;E). S: sample; R: cortical population response; E: estimation report. Dashed vertical lines indicate events of cue display/choice. The only clear (and statistically significant after FDR correction; Figure 3) effect was evident in II(S;R;E) during samples 7 and 8, particularly in the Choice condition (similar, but less pronounced in Cue). N ranges between 27 and 29 participants for different panels, contingent on participant inclusion based on the criterion of at least 81 trials (for computation of intersection information) per condition and sample position (Methods).

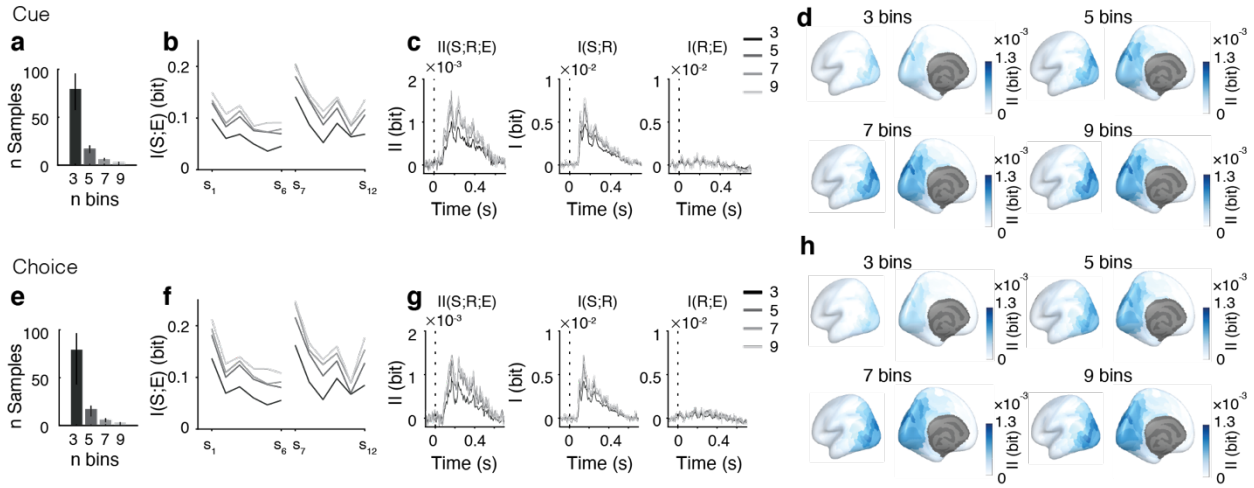

Figure S4. Effect of number of bins on information theoretic measures.

(a) Number of samples per bin of the trivariate probability distribution function for computing intersection information (all trials, 2nd interval sample positions pooled, Cue condition). Error bars indicate the range. **(b)** Information between stimulus (S) and estimation report (E) for different bin numbers. **(c)** Time courses of II(S;R;E) I(S;R), and I(R;E) for different number of bins (dorsal visual cortex). **(d)** Maps of II(S;R;E) on cortical surface for different bin numbers. **(e-h)** As (a-d), but for Choice condition. The qualitative patterns of all information theoretic measures were stable over the range of bin numbers explored, supporting our choice of M=3 to maximize trial counts for the main analyses of the consistency effects on information measures (see Methods). N=30 participants for all panels. Bars, lines; group averages; error bars, SEM.

Supplementary references

1. Murphy, P. R., Wilming, N., Hernandez-Bocanegra, D. C., Prat-Ortega, G. & Donner, T. H. Adaptive circuit dynamics across human cortex during evidence accumulation in changing environments. *Nat Neurosci* **24**, 987–997 (2021).
2. Wilming, N., Murphy, P. R., Meyniel, F. & Donner, T. H. Large-scale dynamics of perceptual decision information across human cortex. *Nat Commun* **11**, 5109 (2020).
3. Waskom, M. L., Okazawa, G. & Kiani, R. Designing and Interpreting Psychophysical Investigations of Cognition. *Neuron* **104**, 100–112 (2019).
4. Kuan, A. T. *et al.* Synaptic wiring motifs in posterior parietal cortex support decision-making. *Nature* **627**, 367–373 (2024).