Interacting with volatile environments stabilizes hidden-state inference and its brain signatures

Aurélien Weiss[1,2,3], Valérian Chambon[2,4], Jan Drugowitsch[5], and Valentin Wyart[1,2,*]

[1] *Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale (Inserm)*

[2] *Département d'Études Cognitives, École Normale Supérieure, PSL University, Paris, France*

[3] *Université Paris Descartes, Paris, France*

[4] *Institut Jean Nicod, Centre National de la Recherche Scientifique (CNRS)*

[5] *Department of Neurobiology, Harvard Medical School, Boston, MA, USA*

* Corresponding author: valentin.wyart@ens.fr

Summary (138 words)

Making accurate decisions in uncertain environments requires identifying the generative cause of sensory cues, but also the expected outcomes of possible actions. Although both cognitive processes can be formalized as Bayesian inference, they are commonly studied using different experimental frameworks, making their formal comparison difficult. Here, by framing a reversal learning task either as cue-based or outcome-based inference, we found that humans perceive the same volatile environment as more stable when inferring its hidden state by interaction with uncertain outcomes than by observation of equally uncertain cues. Time-resolved analyses of magnetoencephalographic (MEG) signals explained this behavioral difference by the neural interaction between inferred beliefs and incoming evidence, an effect originating from the medial temporal lobe (MTL). Together, these findings indicate that the degree of control over the sampling of volatile environments shapes human learning and decision-making under uncertainty.

Introduction (519 words)

Making accurate decisions in an uncertain environment requires inferring its properties from imperfect information (Oaksford and Chater, 2007; Pouget et al., 2013). For example, the decision of eating an apple involves categorizing its variety based on visual *cues*, but also evaluating its expected flavor based on the *outcomes* of similar decisions made in the past. These two types of inference differ qualitatively in the information on which they are based, as in the example above: the variety of an apple is inferred from sensory signals (Shadlen and Kiani, 2013), whereas its expected flavor is retrieved from long-term memory (Shadlen and Shohamy, 2016). These differences make it difficult to formally compare cue-based and outcome-based inference, even though their key difference lies in the degree of control over information sampling rather than the nature of the information itself. Indeed, cue-based inference relies on the presentation of information to an *observer* interpreting its environment, whereas outcome-based inference relies on the active sampling of information by an *agent* interacting with its environment to achieve a particular goal.

It thus remains unclear whether humans learn and decide differently based on the same uncertain information when the information in question corresponds either to external cues, or to outcomes of a previous decision. A formal comparison between cue-based and outcome-based inference requires a shared computational framework that describes them. Yet, canonical 'sequential-sampling' models of sensory evidence accumulation are cast in terms of a continuous random walk process spanning hundreds of milliseconds (Ratcliff and Smith, 2004; Bogacz et al., 2006), whereas 'reinforcement learning' models of action valuation rely on discrete updates of expected outcomes over much longer time scales (Sutton and Barto, 1998, 2018). Another challenge for a direct comparison between the two types of inference comes from the large differences in the experimental paradigms developed to study perceptual (cue-based) decisions and reward-guided (outcome-based) decisions. In particular, perceptual tasks increase uncertainty by decreasing the signal-to-noise ratio of presented stimuli – e.g., the motion coherence of random-dot kinematograms (Gold and Shadlen, 2007; Hanks and Summerfield, 2017). By contrast, reward-guided tasks increase uncertainty by decreasing the predictability of action-outcome contingencies – e.g., differences in reward probability associated with possible actions (Rangel et al., 2008; Rushworth et al., 2011).

Finally, the tasks developed to compare active and passive sampling of novel environments confer large intrinsic benefits to active sampling in terms of information seeking (Gureckis and Markant, 2012; Markant and Gureckis, 2014), thereby rendering comparisons between cue-based and outcome-based inference difficult.

To overcome these challenges, we designed and tested a novel task based on reversal learning (Izquierdo et al., 2017) and a computational framework based on Bayesian inference (Glaze et al., 2015; Wyart and Koechlin, 2016; Drugowitsch et al., 2016), in which cue-based and outcome-based inference can be framed and compared in tightly matched conditions. We recorded magnetoencephalographic (MEG) signals to identify the neural representations and dynamics supporting the two types of inference and their differences. We obtained converging behavioral and neural evidence that interacting with uncertain information stabilizes hidden-state inference, as if humans perceive volatile environments as more stable when interacting with uncertain outcomes than when observing equally uncertain cues.

Results (3,664 words)

Reversal learning task

Healthy adult participants (N = 24) performed a reversal learning task based on visual stimuli which we framed either as cue-based or outcome-based inference, in two conditions corresponding to different blocks of trials. In both conditions, participants were asked to track a hidden state of the task which alternates occasionally and unpredictably between two discrete values (Figure 1). Stimuli corresponded to oriented bars drawn from one of two overlapping probability distributions (categories) centered on orthogonal orientations, each associated with a color (Figure 1A). Each trial consisted of a sequence of 2-8 stimuli after which participants provided a response regarding the hidden state being tracked (Figure 1B). In the cue-based (Cb) condition, participants were instructed to monitor the category (A or B) from which the visual stimuli are drawn (Figure 1C). In the outcome-based (Ob) condition, the same participants were instructed to select the action (left or right key press) which draws stimuli from a target category (counterbalanced across blocks). As indicated above, the hidden state (the drawn category in the cue-based condition, or the target-drawing action in the outcome-based condition) reversed occasionally and unpredictably between trials – thereby requiring participants to adapt their behavior following each reversal.

Importantly, this experimental design allowed for an exact match of all task parameters (including stimulus characteristics, response parameters, presentation times) and all computational variables (predicted by Bayesian inference) between cue-based and outcome-based conditions. In particular, the amount of information provided by each stimulus sequence was strictly identical between conditions (see Methods for analytical derivations).

Slower reversal learning during outcome-based inference

Neither response accuracy – i.e., the fraction of responses consistent with the hidden state (Cb: 81.7 ± 0.7%, Ob: 81.9 ± 0.8%, mean ± s.e.m., paired t-test, t_{23} = 0.3, p = 0.784), nor mean response times (Cb: 594.3 ± 51.6 ms, Ob: 585.3 ± 55.5 ms, t_{23} = -0.5, p = 0.648) differed significantly across conditions. Despite this match, we found a selective difference in reversal learning between cue-based and outcome-based conditions (Figure 2A). Fitting response reversal curves by a saturating

Figure 1 *(legend on next page)*

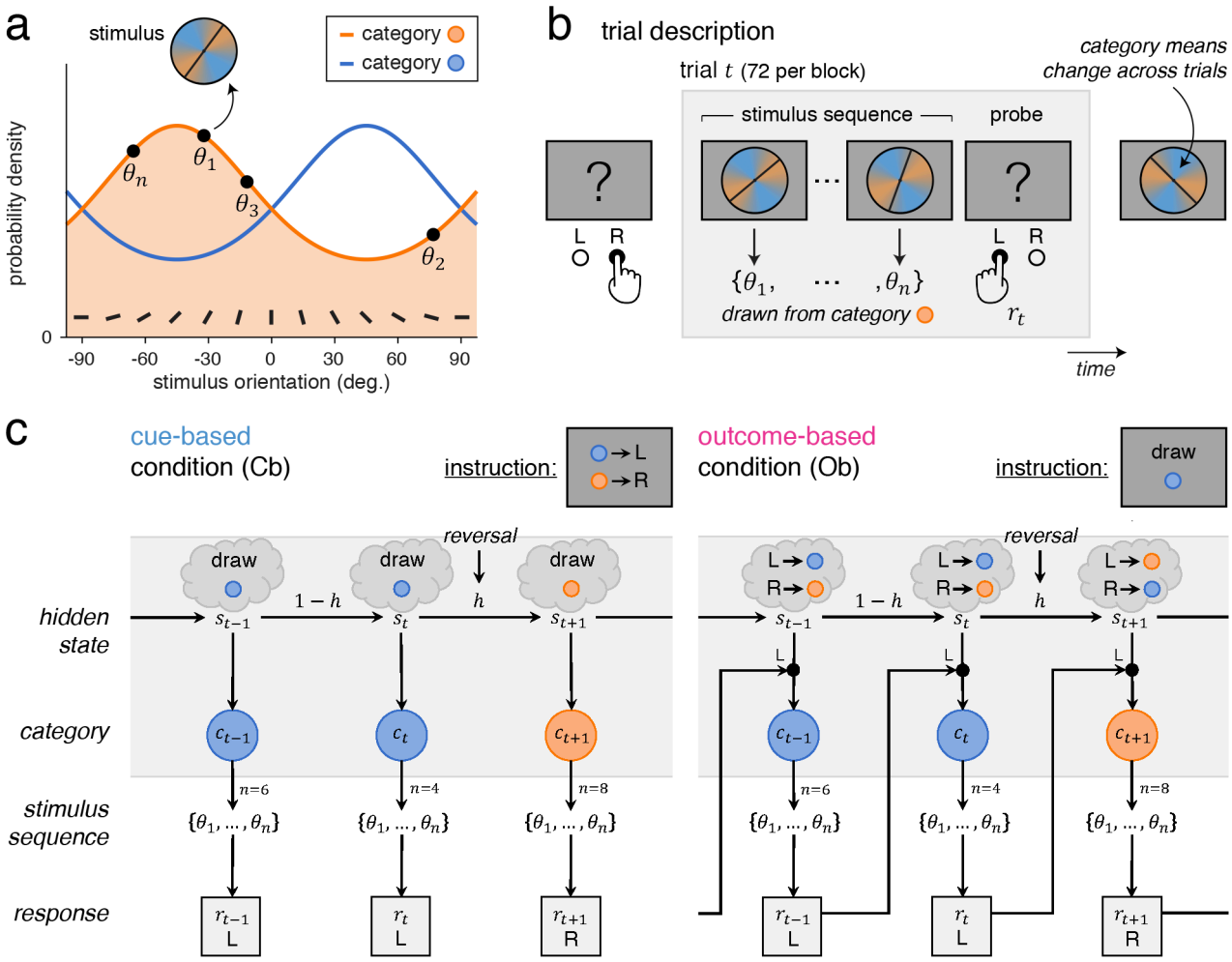

Figure 1 *(figure on previous page)*

Figure 1 – Reversal learning task and conditions. *(**a**) Generative stimulus distributions used for the two categories A and B. Stimuli correspond to bars of orientations* $\{\theta_1, \dots, \theta_n\}$ *drawn from one of two color-labeled categories, associated with overlapping probability density functions centered on orthogonal orientations. (**b**) Trial description. Each trial* t *consists of a sequence of 2-8 stimuli* $\{\theta_1, \dots, \theta_n\}$ *drawn from one of the two categories, presented at an average rate of 2 Hz, after which participants provide a response* r_t *(left or right key press) regarding the hidden state of the task being tracked. (**c**) Graphical description of cue-based and outcome-based conditions. Left: in the cue-based (Cb) condition, participants are instructed to monitor the category (A or B) from which oriented stimuli are drawn. This hidden state* s *of the task alternates occasionally and unpredictably between trials. In this example, the hidden state* s *reverses after trial* t*. Right: in the outcome-based (Ob) condition, the same participants are instructed to select the action (left or right key press) which draws stimuli from a target category. The category drawn at trial* t *depends not only on the hidden state* s_t*, but also on the response provided at the previous trial* r_{t-1}*. As in the cue-based condition, the hidden state* s *(the target-drawing action) alternates occasionally and unpredictably between trials. In this example, the hidden state* s *reverses after trial* t*.*

exponential function revealed a longer reversal time constant in the outcome-based condition (Cb: 0.82 ± 0.10, Ob: 1.28 ± 0.12, t_{23} = 8.0, p < 0.001), as well as a higher asymptotic reversal rate (Cb: 86.2 ± 1.2%, Ob: 90.8 ± 1.2%, t_{23} = 5.2, p < 0.001).

To characterize the origin of these learning differences, we constructed response repetition curves – i.e., the probability of repeating a previous response as a function of the evidence provided by the intervening sequence in favor of the previous response (see Methods). Positive evidence indicates evidence consistent with the previous response, whereas negative evidence indicates evidence conflicting with the previous response. Plotting these response repetition curves revealed a clear leftward shift in the outcome-based condition (Figure 2B). Fitting these curves using a sigmoid function showed a selective increase in the point of subjective equivalence (PSE) – i.e., the amount of conflicting evidence required to switch more often than repeat the previous response (Cb: 1.03 ± 0.10, Ob: 1.43 ± 0.09, t_{23} = 7.7, p < 0.001). Neither the slope of response repetition curves – indexing the sensitivity of responses to evidence (Cb: 2.14 ± 0.16, Ob: 2.10 ± 0.17, t_{23} = -0.2, p = 0.846), nor their lower asymptote – reflecting the small fraction of evidence-independent repetitions (Cb: 3.1 ± 0.9%, Ob: 6.8 ± 1.5%, t_{23} = 2.0, p = 0.056), differed significantly between conditions. Together, these results indicate slower reversal learning in the outcome-based condition, caused by a larger amount of conflicting evidence required to reverse the previous response.

Lower perceived hazard rate during outcome-based inference

The tight match between cue-based and outcome-based conditions allowed us to model human behavior in the two conditions by the same Bayesian inference process (Figure 3A), controlled by two parameters: 1. the perceived hazard rate h – i.e., the *subjective* rate of reversals of the hidden state (Glaze et al., 2015), 2. the inference noise σ – i.e., the standard deviation of internal noise in the processing of each stimulus sequence (Drugowitsch et al., 2016). On each trial, the model updates its belief regarding the current value of the hidden state by combining its *prior* belief, before seeing the new sequence of stimuli, with the evidence provided by each stimulus in the new sequence (see Methods). Given that the hidden state s alternates between two possible values s_1 and s_2, beliefs can be expressed as log-odds, $\log(p(s_1)/p(s_2))$: the sign of the log-odds belief indicates whether s_1 or s_2 is more likely, whereas the magnitude of the log-odds belief indicates the strength

Figure 2 *(legend on next page)*

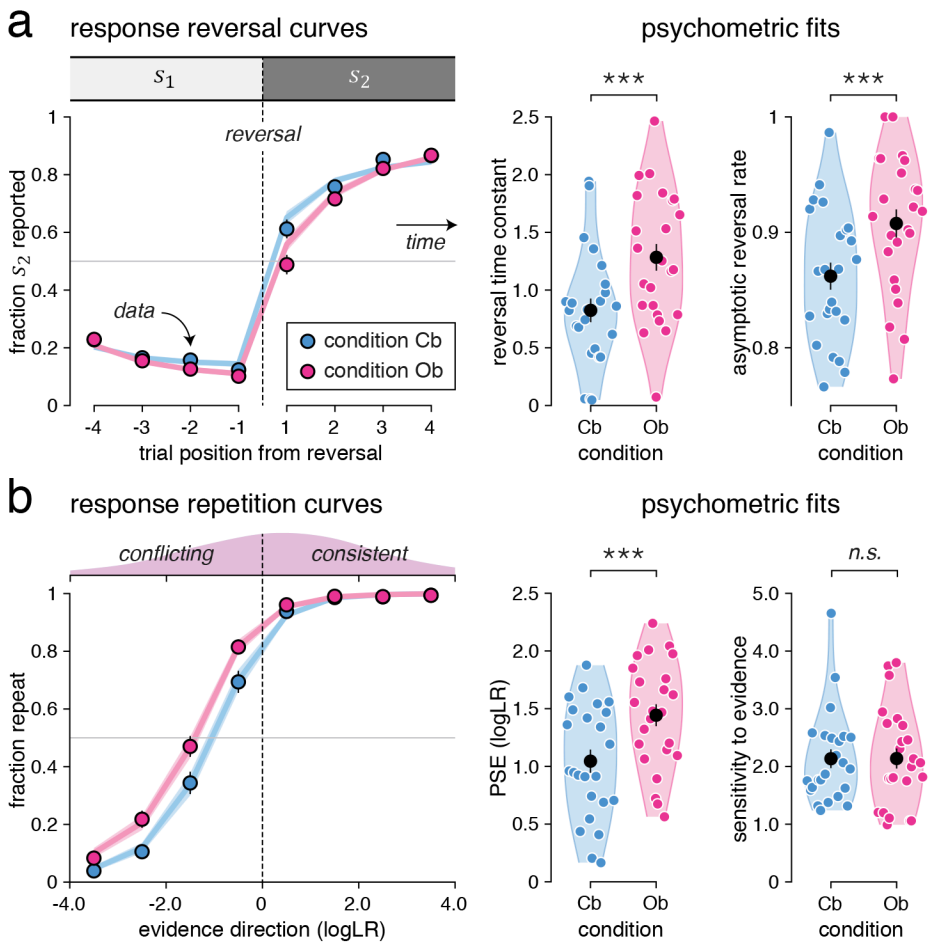

Figure 2 *(figure on previous page)*

Figure 2 – Reversal learning behavior and psychometric fits. *(**a**) Response reversal curves. Left: fraction of hidden state s_2 reported behaviorally (y-axis) in the four trials preceding (left) and following (right) a reversal from s_1 to s_2 (x-axis). The thin dotted line indicates the position of the reversal. Dots indicate the observed data (means ± s.e.m.), whereas lines indicate best-fitting saturating exponential functions. Right: best-fitting parameters of saturating exponential functions in the cue-based and outcome-based conditions. Black dots and error bars indicate group-level means ± s.e.m., whereas colored dots indicate participant-level estimates. The reversal time constant (left) is longer in the outcome-based condition. The asymptotic reversal rate (right) is higher in the outcome-based condition. (**b**) Response repetition curves. Left: fraction of response repetitions (y-axis) as a function of the evidence provided by the intervening sequence in favor of the previous response (x-axis, expressed as logLR). Positive evidence (right) indicates evidence consistent with the previous response, whereas negative evidence (left) indicates evidence conflicting with the previous response. The thin dotted line indicates perfectly uncertain (null) evidence. Same conventions as in panel **a**, except that lines indicate best-fitting sigmoid functions. Right: best-fitting parameters of sigmoid functions in the cue-based and outcome-based conditions. Same conventions as in panel **a**. The PSE (left) is increased in the outcome-based condition, whereas the sensitivity to evidence (right) is equal across conditions. Three stars indicate a significant effect at $p < 0.001$, n.s. a non-significant effect.*

of the belief in favor of the more likely hidden state. Each update of the (log-odds) belief is corrupted by internal noise of standard deviation σ. A fraction of the resulting *posterior* belief is then carried over as prior belief for the next trial as a function of the perceived hazard rate h: the larger the perceived hazard rate, the smaller the prior belief at the beginning of the next trial.

To quantify the (sub)optimality of human performance, we first simulated responses from the Bayesian inference model with optimal parameters (Glaze et al., 2015) – i.e., the true hazard rate (0.125) and exact (noise-free) inference. The response accuracy of the optimal model (88.4 ± 0.3%), identical by construction across the two conditions, substantially exceeded human performance (Cb: t_{23} = 9.3, p < 0.001, Ob: t_{23} = 7.8, p < 0.001). To characterize the origin of human suboptimality and the nature of observed differences between cue-based and outcome-based conditions, we then fitted the perceived hazard rate and the amount of inference noise to human behavior in each condition using particle filtering procedures (see Methods).

Consistent with previous work (Drugowitsch et al., 2016), group-level analyses (Figure 3B) revealed a substantial amount of inference noise which did not differ across conditions (Cb: 0.512 ± 0.024, Ob: 0.550 ± 0.033, t_{23} = 1.5, p = 0.141). Beyond inference noise, we ruled out the presence of a stochastic 'softmax' choice policy in both conditions by Bayesian model selection (Stephan et al., 2009, Cb: exceedance p < 0.001, Ob: exceedance p < 0.001). Together, these results suggest that human suboptimality arises in both conditions from inference noise, whose magnitude correlated significantly between conditions across participants (Pearson r = 0.630, d.f. = 22, p = 0.001). Further group-level analyses (Figure 3C) revealed a lower perceived hazard rate in the outcome-based condition (Cb: 0.191 ± 0.022, Ob: 0.115 ± 0.015, t_{23} = -7.7, p < 0.001), which offers a computational account for the increased PSE in the outcome-based condition. Indeed, the decrease in perceived hazard rate in the outcome-based condition boosts prior beliefs by 42% (Figure 3D), thus requiring more conflicting evidence to reverse a previous response.

This difference in perceived hazard rate between conditions makes a testable prediction: participants should be more accurate in the outcome-based condition in more stable environments where reversals are rare, and more accurate in the cue-based condition in more volatile environments where reversals are frequent (Figure 3E). Unbeknownst to participants, we varied the true

Figure 3 *(legend on next page)*

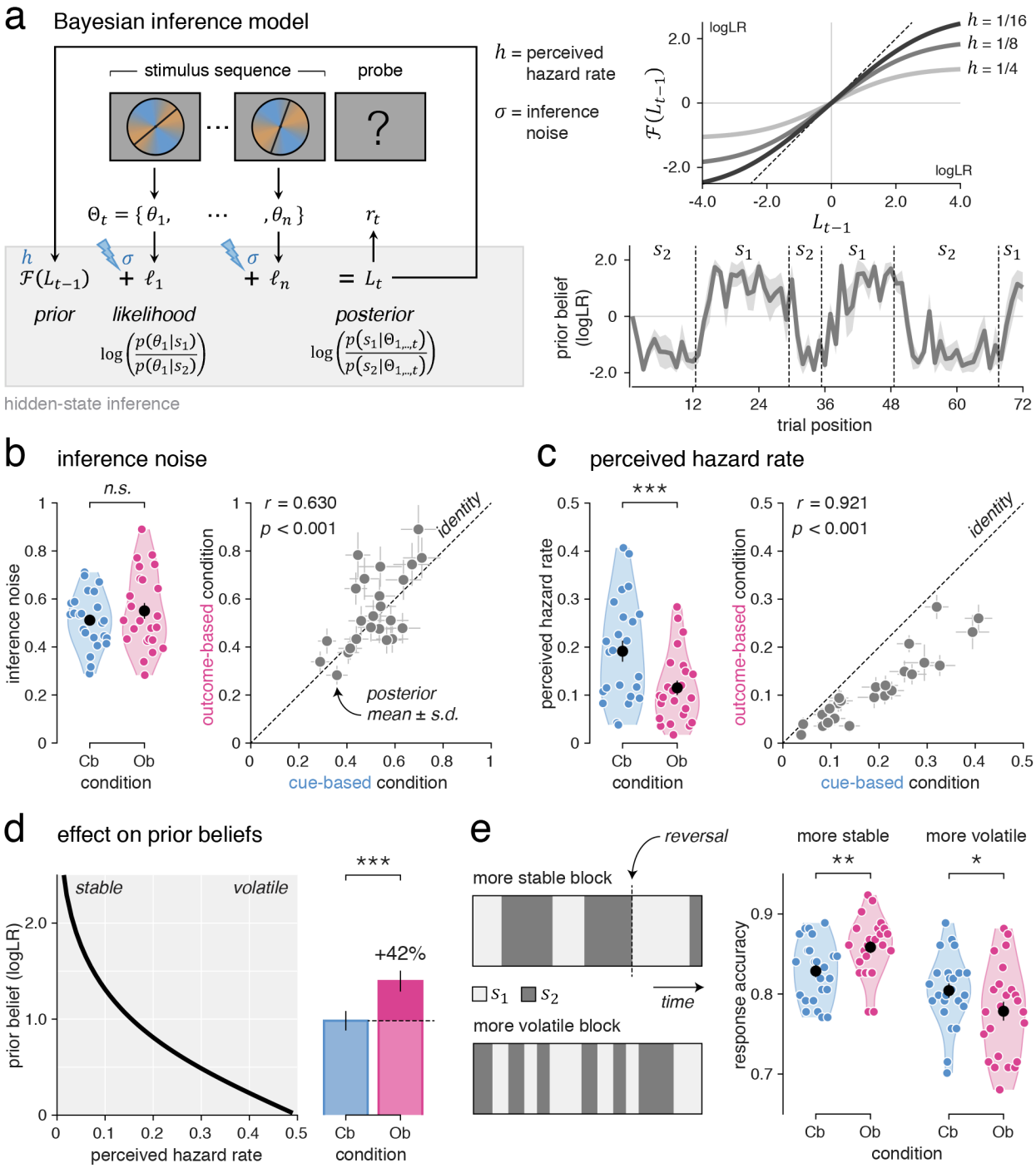

Figure 3 *(figure on previous page)*

Figure 3 – Computational modeling of behavior. *(**a**) Left: graphical description of the Bayesian inference model, controlled by two scalar parameters: 1. the perceived hazard rate h (the subjective rate of reversals of the hidden state), and 2. the inference noise σ (the standard deviation of internal noise in the processing of each stimulus sequence). On each trial, the model updates its belief L regarding the current value of the hidden state s_t by combining its prior belief $\mathcal{F}(L_{t-1})$ with the likelihood $\mathcal{L}_t \equiv \{\ell_1, \dots, \ell_n\}$ associated with the stimuli in the new sequence, parameterized by their orientations $\Theta_t \equiv \{\theta_1, \dots, \theta_n\}$, resulting in the posterior belief L_t. Right top: modelled fraction of the posterior belief at the end of trial $t-1$, L_{t-1} (x-axis), carried over as prior belief at the beginning of trial t, $\mathcal{F}(L_{t-1})$ (y-axis), for three values of perceived hazard rate (gray-shaded lines). The larger the perceived hazard rate, the smaller the prior belief at the beginning of the next trial. Right bottom: simulated trajectory of the prior belief at the beginning of each trial over the course of an example block with $h = 1/8$ and $\sigma = 0.5$. The line and shaded error bar indicate the mean and s.d. across simulations. Vertical dotted lines indicate reversals of the hidden state. (**b**) Inference noise estimates. Left: inference noise estimates in the cue-based and outcome-based conditions. Same conventions as in **Figure 2**. Right: correlation between inference noise estimates in cue-based (x-axis) and outcome-based (y-axis) conditions. Dots and error bars indicate posterior means ± s.d. obtained by model fitting. The thin dotted line shows the identity line. Inference noise estimates do not differ between conditions. (**c**) Perceived hazard rate estimates. Left: perceived hazard rate estimates in the cue-based and outcome-based conditions. Same conventions as in panel **b**. Right: correlation between perceived hazard rate estimates in cue-based and outcome-based conditions. Same conventions as in panel **b**. The perceived hazard rate is significantly lower in the outcome-based condition. (**d**). Left: predicted effect of perceived hazard rate (x-axis) on the magnitude of prior beliefs (y-axis). Dots and error bars indicate means ± s.e.m. Right: predicted magnitude of prior beliefs in the cue-based and outcome-based conditions. Bars and error bars indicate means ± s.e.m. (**e**) Effect of volatility on the behavioral difference between conditions. Left: evolution of the hidden state s over the course of a more stable block (top) and a more volatile block (bottom). Right: response accuracy is higher for the outcome-based condition in more stable blocks (left), but worse in more volatile blocks (right). One star indicates a significant effect at $p < 0.05$, two stars at $p < 0.01$, three stars at $p < 0.001$, n.s. a non-significant effect.*

hazard rate across blocks between 0.083 (more stable) and 0.167 (more volatile). As predicted, participants were more accurate in the outcome-based condition in more stable blocks (t_{23} = 3.4, p = 0.002), and more accurate in the cue-based condition in more volatile blocks (t_{23} = -2.5, p = 0.020, interaction: $F_{1,23}$ = 20.2, p < 0.001). This pattern supports the conclusion that participants perceive the outcome-based condition as more stable than the cue-based condition, despite the same true hazard rate used in the two conditions.

Neural correlates of stimulus processing across space and time

To identify the neural representations and dynamics supporting cue-based and outcome-based inference, we described each stimulus (1,440 per condition and per participant) by a set of distinct characteristics: its orientation, its change (tilt) from the preceding stimulus, and the strength of the evidence it provides to the inference process. We then applied multivariate pattern analyses to MEG signals (Grootswagers et al., 2017) aligned to stimulus onset to estimate the neural patterns, or 'codes', associated with these characteristics. Due to the fine temporal resolution of MEG signals (Baillet, 2017), we could extract the time course of neural information processing within the first hundreds of milliseconds following stimulus onset (see Methods).

The neural coding of stimulus orientation (Figure 4A) peaked at 120 ms following stimulus onset (jackknifed mean, Cb: 123.1 ms, Ob: 122.4 ms), with equal precision across cue-based and outcome-based conditions. The neural coding of stimulus change (Figure 4B) – defined as the absolute tilt between the current stimulus and its predecessor in the sequence – peaked at 220 ms following stimulus onset (Cb: 225.7 ms, Ob: 223.3 ms), again with equal precision across conditions. The neural coding of stimulus evidence (Figure 4C) – defined as the absolute tilt between the current stimulus and the nearest category boundary – peaked around 360 ms following stimulus onset (Cb: 357.5 ms, Ob: 378.7 ms). Unlike the two previous characteristics, the computation of stimulus evidence requires the mapping of stimulus orientation onto category-defined axes, which varied randomly across trials. Despite its relevance for inference, the coding precision of stimulus evidence did not differ between cue-based and outcome-based conditions.

To assess the degree of similarity of these neural codes between conditions, we performed cross-condition generalization analyses: we used the coding weights estimated in one condition to

Figure 4 *(legend on next page)*

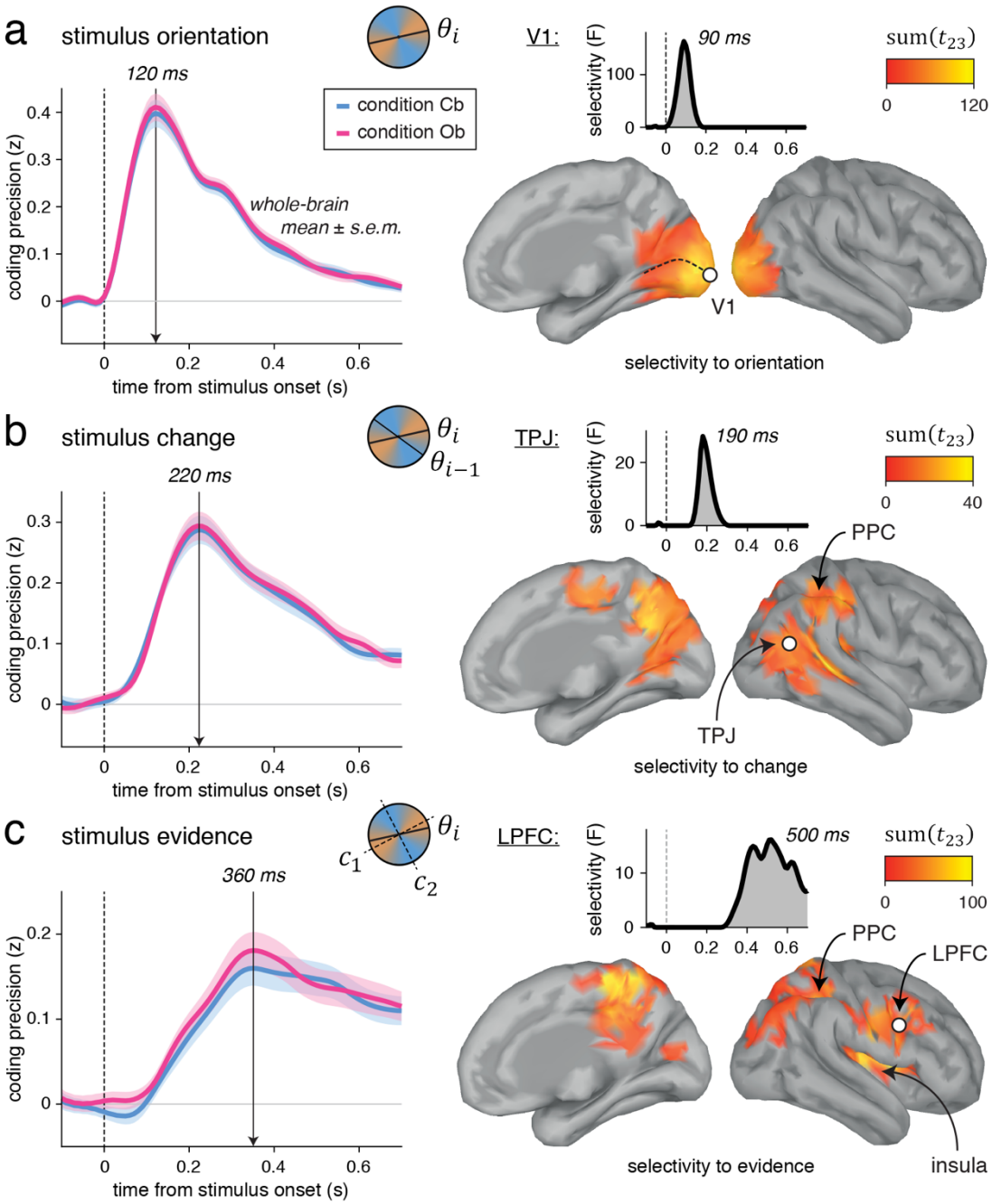

Figure 4 *(figure on previous page)*

Figure 4 – Neural correlates of stimulus processing. *(**a**) Neural coding of stimulus orientation. Left: time course of neural coding from whole-brain MEG sensors. Stimulus orientation coding peaks at 120 ms following stimulus onset, with equal precision across conditions. Lines and shaded error bars indicate means ± s.e.m. Right: searchlight-based coding of stimulus orientation across the cortical surface (left: medial view, right: lateral view). The bilateral spatiotemporal cluster found in early visual cortex (including V1), marginalized across time (cluster definition threshold $p = 0.001$, cluster-level $p < 0.001$) corresponds to cortical sources for which the coding precision of stimulus orientation significantly exceeds the coding precision of stimulus change and stimulus evidence. The thin dotted line follows the calcarine sulcus. Inset: time course of the selectivity statistic for a cortical source in V1 (white dot). (**b**) Neural coding of stimulus change. Left: time course of neural coding. Stimulus change coding peaks at 220 ms following stimulus onset, with equal precision across conditions. Same conventions as in panel **a**. Right: searchlight-based coding of stimulus change across the cortical surface. Change-selective clusters are found in the TPJ and the PPC. Inset: time course of the selectivity statistic for a cortical source in the TPJ (white dot). (**c**) Neural coding of stimulus evidence. Left: time course of neural coding. Stimulus evidence peaks at 360 ms following stimulus onset, with equal precision across conditions. Same conventions as in panels **a** and **b**. Right: searchlight-based coding of stimulus evidence across the cortical surface. Evidence-selective clusters are found in the PPC, the LPFC and the insula. Inset: time course of the selectivity statistic for cortical source in the LPFC (white dot).*

compute neural predictions in the other condition (see Methods). Furthermore, we devised a procedure for estimating the degree of similarity between two neural codes based on a population of linear coding units (Figure 5A). Applying this procedure to the neural codes of each stimulus characteristic indicated near-perfect similarity between cue-based and outcome-based conditions – i.e., no loss in coding precision when using coding weights estimated in one condition to compute neural predictions in the other condition (Figure 5B). These results suggest that stimulus processing in the cue-based and outcome-based conditions relies on shared neural processes.

Next, we sought to identify the cortical distribution of these shared neural codes across conditions. For this purpose, we estimated the cortical sources of observed MEG signals and performed focal multivariate pattern analyses in source space using a 'searchlight' approach (see Methods). To increase the spatial selectivity of obtained results, we focused on cortical sources for which the coding precision of one stimulus characteristic significantly exceeds the coding precision of the other two characteristics. This analysis identified a cluster in early visual cortex for stimulus orientation, peaking at 90 ms following stimulus onset (Figure 4A). For stimulus change, the same analysis isolated clusters in the temporo-parietal junction (TPJ) and the posterior parietal cortex (PPC), peaking at 190 ms following stimulus onset (Figure 4B). Finally, the coding of stimulus evidence was associated with clusters in the PPC and the lateral prefrontal cortex (LPFC) after 400 ms following stimulus onset (Figure 4C). Together, these results highlight a spatiotemporal cascade of neural patterns involved in the processing of each stimulus, shared across the two conditions.

Neural correlates of the interaction between evidence and beliefs

On each trial, inference requires updating the prior belief in the current value of the hidden state (the drawn category in the cue-based condition, or the target-drawing action in the outcome-based condition) with the evidence provided by each stimulus in the new sequence (Figure 3A). To isolate the neural patterns involved in the interaction between evidence and beliefs, we asked how well we could decode from MEG signals the evidence provided by each stimulus in favor of the previous response – which reflects the prior belief at the beginning of the current trial. Positive evidence indicates evidence consistent with the previous response, whereas negative evidence indicates evidence conflicting with the previous response.

Figure 5 *(legend on next page)*

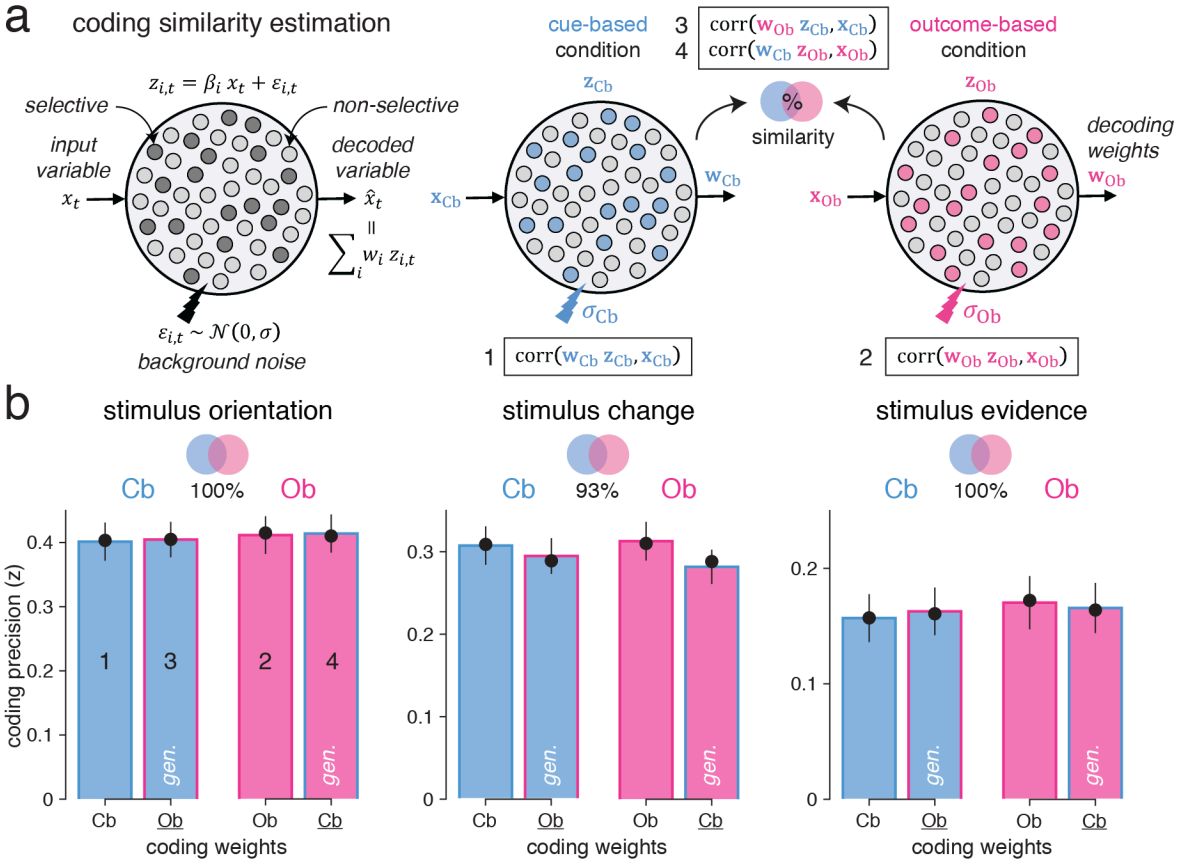

Figure 5 *(figure on previous page)*

Figure 5 – Estimation of neural coding similarity across conditions. *(**a**) Graphical description of the coding similarity estimation procedure. Left: a population of linear coding units $z_{i,t}$ represents an input scalar variable x_t by a fraction of 'selective' units ($z_{i,t} = \beta_i\, x_t + \varepsilon_{i,t}$ where $\beta_i \neq 0$) with background noise $\varepsilon_{i,t}$ of standard deviation σ. A linear decoder is applied to compute an estimate $\hat{x}_t$ of the input variable from population activity. Right: the same population of coding units represents the same input variable x by partially overlapping selective units z in the cue-based (left) and outcome-based (right) conditions. Computing coding precision within each condition (1-2) and across conditions (3-4, marked by gen., by using the coding weights w estimated in one condition to compute neural predictions in the other condition) allows to quantify the degree of similarity (overlap) between selective units in the two conditions. (**b**) Estimated coding similarities for stimulus orientation (left), stimulus change (middle) and stimulus evidence (right). Bars and error bars indicate means ± s.e.m. Dots show predicted values obtained by simulating the population of coding units with best-fitting estimates of similarity and background noise. Bar fillings indicate the condition in which neural predictions $\hat{\mathrm{x}}$ are computed. Bar outlines indicate the condition in which coding weights w are estimated. This procedure indicates near-perfect coding similarity between cue-based and outcome-based conditions for stimulus orientation (left, jackknifed mean: 100%), stimulus change (center, 93%) and stimulus evidence (right, 100%).*

We could decode this 'consistency' variable in both conditions over a large time window starting around 110 ms following stimulus onset (Figure 6A). Furthermore, and in contrast to other stimulus characteristics which do not depend on participants' beliefs, we could decode this variable with higher precision in the outcome-based condition over the entire time window (Cb: 0.076 ± 0.017, Ob: 0.142 ± 0.017, peak t_{23} = 6.6, cluster-level $p < 0.001$). The cross-condition generalization analysis of this neural code revealed near-perfect similarity across conditions (Figure 6B): we could decode consistency in the cue-based condition marginally better when using coding weights estimated in the outcome-based condition (t_{23} = 2.1, p = 0.051) where coding precision is substantially higher. This effect could be nullified with as little as 20% dissimilarity across conditions. These results indicate that consistency is coded by the same neural patterns in the two conditions, only with higher precision in the outcome-based condition.

Source-level analyses identified two clusters responsible for the increased coding precision of consistency in the outcome-based condition: one in the right medial temporal lobe (MTL) and another in the PPC (Figure 6C, cluster-level $p < 0.001$). The coding precision of other stimulus characteristics, including stimulus evidence, did not differ between conditions in either cluster. The two clusters differed in their timings of effect (Figure 6D), with the MTL showing an early difference between conditions (50% of the cluster mass between 180 and 280 ms following stimulus onset), and the PPC showing a later difference (50% of the cluster mass between 380 and 480 ms). These findings indicate that the neural interaction between incoming evidence and prior beliefs – beliefs which are known to be boosted in the outcome-based condition (Figure 3D), is coded with higher precision in the MTL and PPC in the outcome-based condition.

Slower reversal of neural beliefs during outcome-based inference

The neural coding of the interaction (consistency) between incoming evidence and prior beliefs offers a handle to explain behavioral differences between conditions. Behavioral analyses have revealed that participants perceive the outcome-based condition as more stable than the cue-based condition, and require more conflicting evidence to reverse a previous response. We reasoned that the boosted prior beliefs in the outcome-based condition could explain the increased neural coding of the interaction between incoming evidence and prior beliefs.

Figure 6 *(legend on next page)*

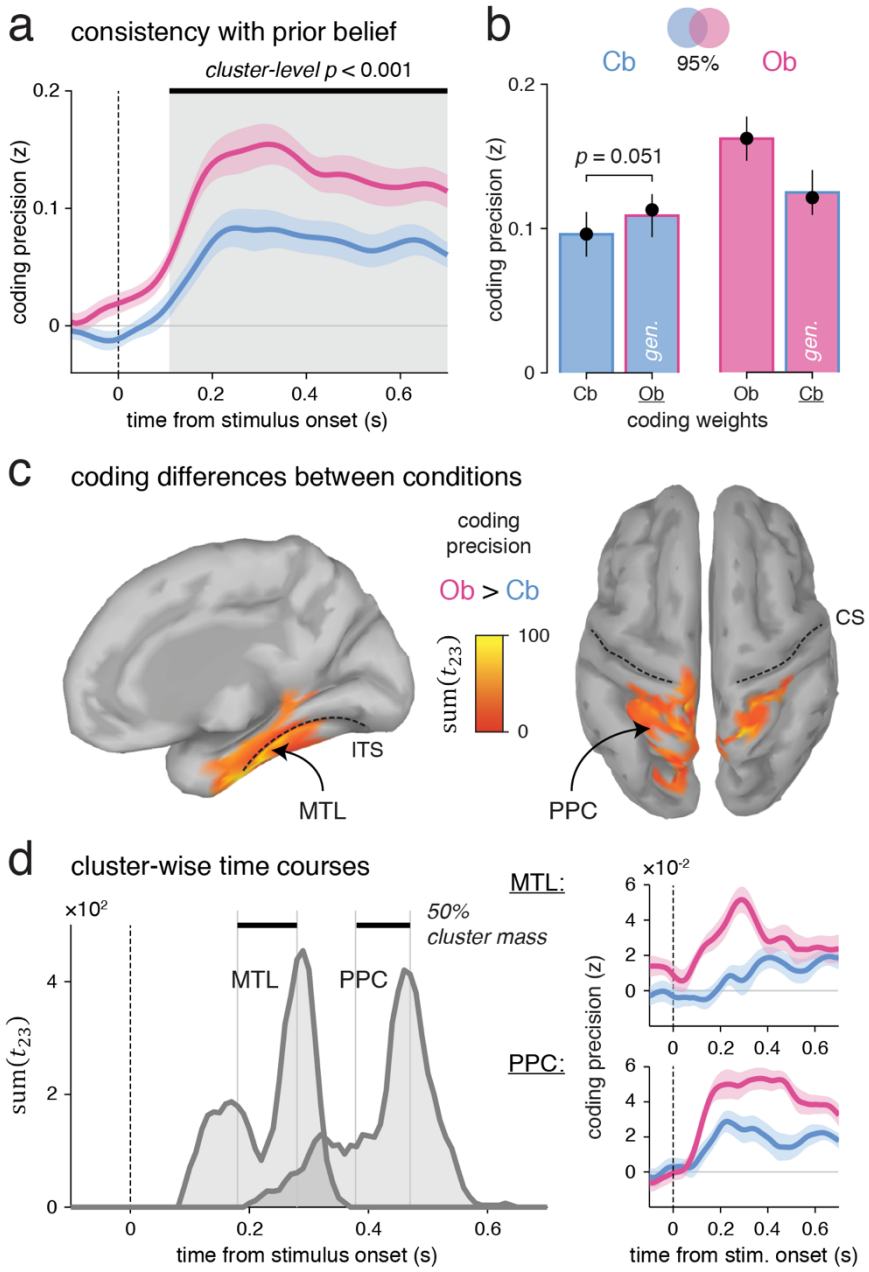

Figure 6 *(figure on previous page)*

Figure 6 – Neural correlates of the interaction between evidence and beliefs. *(**a**) Neural coding of stimulus-wise consistency with the prior belief, reflected in the previous response. Consistency is coded with higher precision in the outcome-based condition. The shaded area indicates the significant difference in coding precision between conditions. Lines and error bars indicate means ± s.e.m. (**b**) Estimated coding similarity for consistency across conditions. Cross-condition generalization indicates near-perfect similarity across conditions (jackknifed mean: 95%). Bars and error bars indicate means ± s.e.m. Dots show best-fitting values from the similarity estimation procedure. Same color conventions as in **Figure 5**. (**c**) Whole-brain differences in coding precision between conditions, marginalized across time (cluster definition threshold p = 0.001, cluster-level p < 0.001). Source-level analyses identify one spatiotemporal cluster in the right MTL (left) and another in the PPC (right). Thin dotted lines follow the inferotemporal sulcus (ITS, left) and the central sulcus (CS, right). (**d**) Cluster-wise time courses of the difference in coding precision for the MTL and PPC clusters. Left: cluster-wise statistics marginalized across space. The MTL cluster shows an early (180-280 ms) difference, whereas the PPC shows a later (380-480 ms) difference. Horizontal black lines indicate the latencies of 50% cluster mass. Right: neural coding of consistency with the prior belief for the MTL cluster (top) and the PPC cluster (bottom).*

Until now, we defined consistency relative to the prior belief at the beginning of the current trial, before the new sequence of stimuli. However, the belief can switch in the middle of a sequence due to conflicting evidence – e.g., in the first trials following a reversal of the hidden state. These 'mid-sequence' belief switches are less likely to occur when prior beliefs are of larger magnitude because more conflicting evidence is required to trigger them, as in the outcome-based condition. If consistency is coded not in relation to the *prior* belief at the beginning of the current trial, but the *current* belief accounting for the evidence provided by previous stimuli in the current sequence, then mid-sequence belief switches should degrade the neural coding of consistency defined relative to the prior belief – precisely because the prior belief does not match the current belief following mid-sequence switches.

To test whether mid-sequence belief switches could explain the lower coding precision of consistency in the cue-based condition where they are more numerous, we redefined consistency in relation to the current belief accounting for previous stimuli in the same sequence. For this purpose, we estimated in each condition the belief trajectory over the course of each sequence based on: 1. the evidence provided by each stimulus in the sequence, and 2. best-fitting values for the perceived hazard rate h and the amount of inference noise σ (see Methods). Estimated belief trajectories thus accounted for the lower perceived hazard rate, and the larger prior beliefs, found in the outcome-based condition (Figure 7A).

As predicted, we could decode consistency from MEG signals more precisely when defined relative to the current belief than the prior belief, from 150 ms following stimulus onset (Figure 7B, peak $F_{1,23} = 28.2$, cluster-level $p < 0.001$). Furthermore, accounting for the difference in perceived hazard rate in the definition of consistency explained away most of the coding difference between conditions (interaction between 270 and 470 ms, peak $F_{1,23} = 23.7$, cluster-level $p < 0.001$). This pattern of findings suggests that the increased neural coding of consistency in the outcome-based condition reflects the lower perceived hazard rate in this condition. To further support this claim, we estimated in each condition and for each participant the 'neural' hazard rate for which we could best decode consistency from MEG signals – i.e., ignoring the best-fitting value for the perceived hazard rate obtained from behavior (Figure 7C). This analysis revealed a lower neural hazard rate in the

outcome-based condition (Cb: 0.319 ± 0.023, Ob: 0.216 ± 0.026, t_{23} = -4.5, p < 0.001), thereby mirroring the lower perceived hazard rate found in this condition.

We conducted an additional analysis to estimate belief trajectories directly from the neural coding of consistency – with the idea of visualizing the larger prior beliefs in the outcome-based condition. We reasoned that the neural coding of consistency, defined in relation to the belief at the beginning of the current trial, should switch sign when the belief itself switches in the middle of a sequence. These mid-sequence belief switches are bound to occur in trials leading to response reversals (where the belief at the end of the trial has opposite sign to the belief at the beginning of the trial), but not in sequences leading to response repetitions (where the belief at the end of the trial has the same sign as the belief at the beginning of the trial). Furthermore, these coding switches should occur later inside a sequence in the outcome-based condition where prior beliefs are larger and thus require more conflicting evidence to change sign. To test these predictions, we decoded consistency at each stimulus position in a sequence (Figure 7D), separately for sequences leading to response reversals and sequences leading to response repetitions.

Our predictions were corroborated by the neural data. First, sequences leading to response reversals were associated with coding switches: the first stimulus in the sequence was coded positively – i.e., in direction of the belief at the beginning of the trial (Cb: 0.049 ± 0.018, t_{23} = 2.7, p = 0.011, Ob: 0.111 ± 0.026, t_{23} = 4.3, p < 0.001), whereas the last stimulus in the same sequence was coded negatively – i.e., in direction of the belief at the end of the trial (Cb: -0.124 ± 0.016, t_{23} = -7.7, p < 0.001, Ob: -0.111 ± 0.019, t_{23} = -5.9, p < 0.001). These coding switches were not observed during sequences leading to response repetitions: the last stimulus was also coded positively (Cb: 0.172 ± 0.018, t_{23} = 9.6, p < 0.001, Ob: 0.185 ± 0.016, t_{23} = 11.5, p < 0.001). Second, coding switches occurred significantly later in the outcome-based condition (jackknifed regression of zero-crossing times, t_{23} = 4.6, p < 0.001). Together, these findings support the presence of larger prior beliefs in the outcome-based condition.

Neural dissociation between absolute and relational coding of evidence

Neural responses to each stimulus revealed two concurrent coding schemes of the evidence provided to the inference process (Figure 8A): 1. an ‘absolute’ coding reflected in the strength of the

Figure 7 *(legend on next page)*

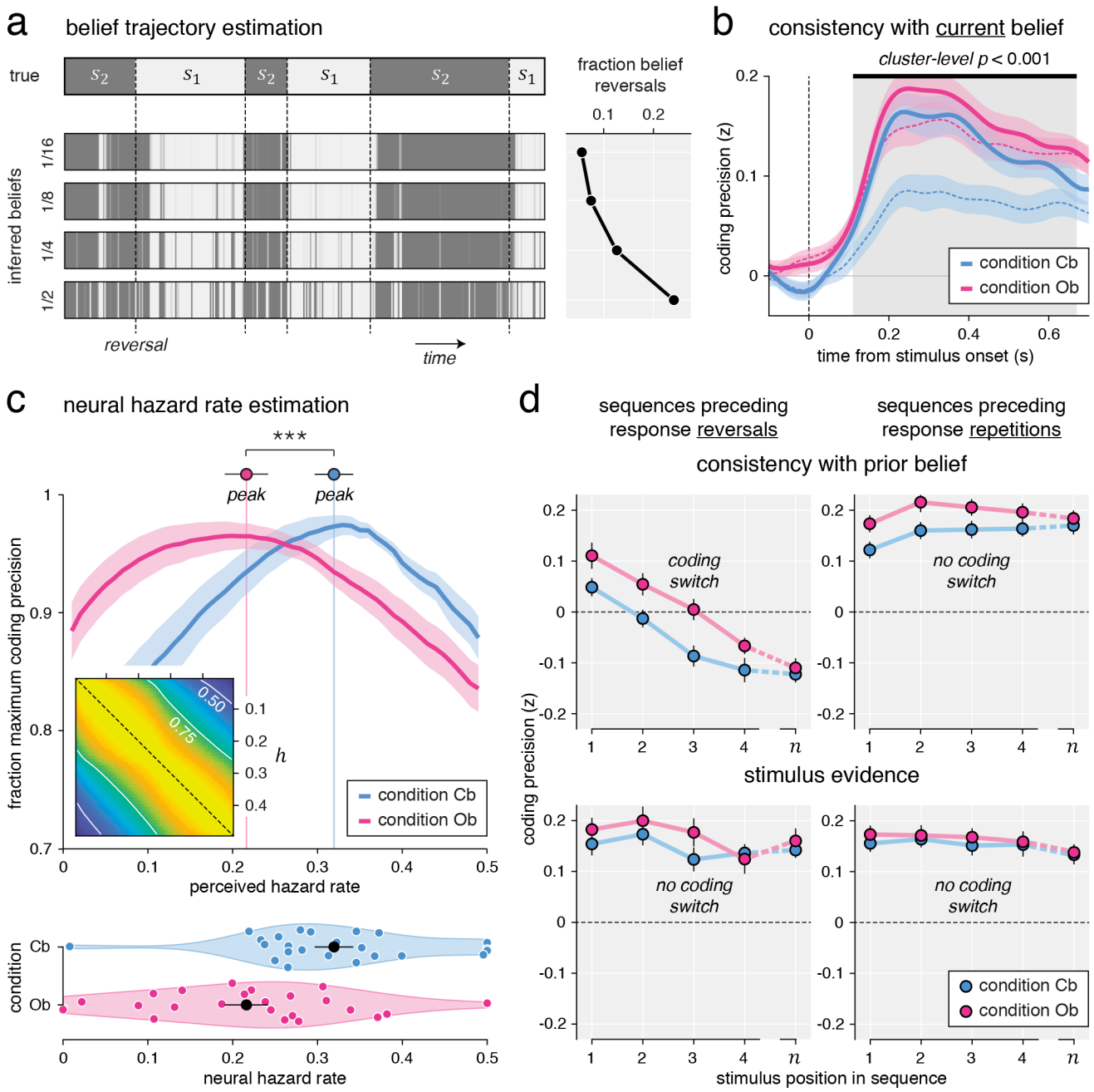

Figure 7 *(figure on previous page)*

Figure 7 – Belief trajectory estimation and analyses. *(**a**) Estimation of belief trajectory from behavior and perceived hazard rate. Left: each horizontal strip indicates the evolution of the hidden state s over the course of an example block. The upper strip indicates the true hidden state (with vertical dotted lines indicating reversals), whereas the lower strips indicate inferred beliefs in the current hidden state for increasing perceived hazard rates, from 1/16 (top) to 1/2 (bottom). Belief trajectories at trial t are conditioned on observed responses up to trial t using particle filtering. Right: fraction of mid-sequence belief reversals associated with each belief trajectory, growing with the perceived hazard rate. (**b**) Neural coding of stimulus-wise consistency with the current belief, computed using the estimated belief trajectory based on the best-fitting perceived hazard rate in each condition (lower in the outcome-based condition). Thin dotted lines show the neural coding of consistency with the prior belief, for comparison with **Figure 6**. Lines and shaded error bars indicate means ± s.e.m. Consistency is coded more precisely in relation to the current belief (thick lines) than to the prior belief (thin dotted lines). (**c**) Estimation of neural hazard rate. Relative coding precision of consistency (y-axis, expressed as the fraction of the maximum per participant and condition) as a function of perceived hazard rate (x-axis). The neural hazard rate is defined as the perceived hazard rate which maximizes coding precision (dots and error bars indicate its mean ± s.e.m.). Lines and shaded error bars indicate means ± s.e.m. Inset: cross-correlation (r-squared) matrix for consistency across perceived hazard rates. (**d**) Neural coding trajectories preceding response reversals (left column) and repetitions (right column), for consistency with the prior belief (top row) and the strength of stimulus evidence (bottom row). Sequences leading to response reversals are associated with switches in the neural coding of consistency, not stimulus evidence, which occur later in the outcome-based condition. Dots and error bars indicate means ± s.e.m. Three stars indicate a significant effect at $p < 0.001$.*

evidence, and 2. a 'relational' coding reflected in the consistency of the same evidence with current beliefs about the hidden state. Examining their time courses (Figure 8B) revealed a marked difference in coding dynamics: the relational (belief-dependent) coding scheme showed a faster rise in precision than the absolute (belief-independent) coding scheme, from 70 to 300 ms following stimulus onset (peak t_{23} = 7.0, cluster-level p < 0.001), and peaked about 120 ms earlier (jackknifed mean, relational: 243.4 ms, absolute: 367.6 ms, t_{23} = -5.6, p < 0.001).

A second fundamental difference between the two coding schemes concerns how their precision varies with the magnitude of beliefs (Figure 8C). Indeed, the precision of relational coding scaled positively with the magnitude of beliefs (linear regression, Cb: β = 0.052 ± 0.008, t_{23} = 6.3, p < 0.001, Ob: β = 0.063 ± 0.010, t_{23} = 6.7, p < 0.001) – as expected from a coding scheme which reflects the consistency of incoming evidence in relation to the current belief (and is thus not defined in the absence of beliefs). By contrast, the precision of absolute coding did not increase but rather decreased slightly with larger beliefs (Cb: β = -0.013 ± 0.005, t_{23} = -2.3, p = 0.031, Ob: -0.017 ± 0.004, t_{23} = -4.3, p < 0.001). In other words, the absolute coding scheme reflects the *objective* (veridical) amount of evidence provided by each stimulus, whereas the relational coding scheme reflects the interaction between the same evidence and *subjective* (inferred) beliefs.

Last, we identified cortical sources for which the precision of relational coding is significantly larger than the precision of absolute coding (Figure 8D). Averaging across conditions revealed a bilateral temporal cluster (cluster-level p < 0.001), which overlapped broadly with the MTL cluster identified earlier when comparing cue-based and outcome-based conditions (Figure 6C). The timing of effect in the temporal cluster (50% of the cluster mass between 200 and 260 ms following stimulus onset) matched the early rise of relational coding. To assess the spatial specificity of this effect, we contrasted relational and absolute coding precision in the MTL and PPC clusters identified earlier by contrasting cue-based and outcome-based conditions (Figure 8E). In the early time window associated with the MTL cluster (between 180 and 280 ms), the MTL showed stronger relational coding (Cb: 0.021 ± 0.008, t_{23} = 2.8, p = 0.011, Ob: 0.045 ± 0.007, t_{23} = 6.3, p < 0.001), an effect more pronounced in the outcome-based condition (t_{23} = 3.0, p = 0.007). By contrast, in the late time window associated with the PPC cluster (between 380 and 480 ms), the PPC showed stronger absolute

coding (Cb: -0.017 ± 0.007, t_{23} = -2.4, p = 0.026, Ob: -0.020 ± 0.008, t_{23} = -2.4, p = 0.022), whereas the MTL did not show any coding difference (Cb: 0.009 ± 0.007, t_{23} = 1.3, p = 0.207, Ob: -0.001 ± 0.011, t_{23} = -0.1, p = 0.916). These results support an active involvement of the MTL in the relational coding of evidence – a coding scheme enhanced during outcome-based inference.

Figure 8 *(legend on next page)*

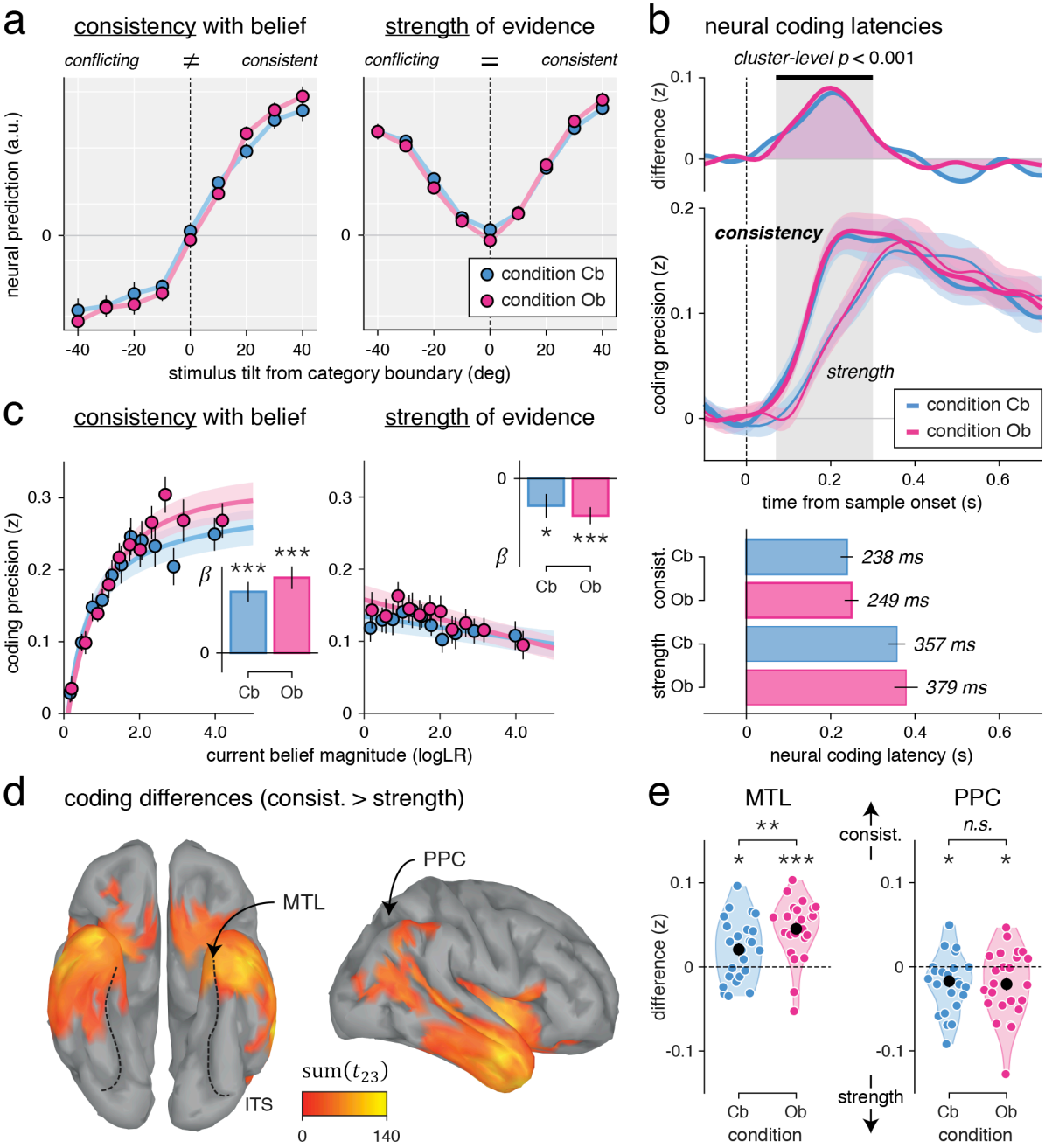

Figure 8 *(figure on previous page)*

Figure 8 – Neural correlates of absolute and relational coding of evidence. *(**a**) Neural predictions for stimulus-wise consistency with the current belief (left) and strength of evidence (right), binned as a function of stimulus tilt from the nearest category boundary. Stimulus tilt is signed in relation to the current belief: positive tilt indicates a stimulus consistent with the current belief, whereas negative tilt indicates a stimulus conflicting with the current belief. Neural predictions of consistency (left) have opposite signs for belief-consistent and belief-conflicting evidence (relational coding scheme), whereas neural predictions of evidence strength (right) are independent of the current belief (absolute coding scheme). Dots and error bars indicate means ± s.e.m. (**b**) Neural coding latencies for consistency and strength of evidence. Top: neural coding time courses (consistency: thick lines; strength: thin lines). The shaded area indicates the significant difference in coding precision between consistency and strength of evidence (shown above the neural coding time courses for each condition). Consistency shows a faster rise in coding precision than the strength of evidence. Lines and shaded error bars indicate means ± s.e.m. Bottom: estimated neural coding latencies. The neural coding of consistency peaks 120 ms earlier than the strength of evidence. Bars and error bars indicate jackknifed means ± s.e.m. (**c**) Effect of current belief magnitude on the neural coding of consistency (left) and strength of evidence (right), grouped in equally sized bins. The coding precision of consistency (left), but not the strength of evidence (right), scales positively with belief magnitude. Dots and error bars indicate means ± s.e.m. Lines and shaded error bars show parametric fits. Inset: linear regression slope estimates (means ± s.e.m) of coding precision against belief magnitude. (**d**) Searchlight-based differences in coding precision between consistency and strength of evidence, averaged across conditions and marginalized across time (cluster definition threshold $p = 0.001$, cluster-level $p < 0.001$). The bilateral temporal cluster overlaps broadly with the MTL cluster identified by comparing conditions. The thin dotted line follows the inferotemporal sulcus (ITS). (**e**) Coding precision differences (consistency minus strength of evidence) in the MTL (left) and PPC (right) clusters identified by comparing conditions. The MTL cluster shows stronger relational coding (consistency > strength), whereas the PPC cluster shows stronger absolute coding (consistency < strength). One star indicates a significant effect at $p < 0.05$, two stars at $p < 0.01$, three stars at $p < 0.001$, n.s. a non-significant effect.*

Discussion (1,683 words)

Accurate decision-making in uncertain environments requires identifying the generative cause of observed stimuli (cue-based inference, as in perceptual decisions), but also the expected consequences of one's own actions (outcome-based inference, as in reward-guided decisions). These two types of inference differ in the degree of control the decision-maker has over the sampling of evidence, a variable that is usually confounded with other task variables in paradigms used to study perceptual and reward-guided decisions. By comparing cue- and outcome-based inference in tightly matched conditions, we show that *interacting* with uncertain evidence – rather than *observing* the same evidence – increases the perceived stability of volatile environments.

A first immediate conclusion is that the status of the decision-maker – an *agent* interacting with its environment in the outcome-based condition, or an *observer* contemplating the same environment in the cue-based condition – shapes human learning and decision-making under uncertainty. Surprisingly, existing theories of inference either do not consider the degree of control over the sampling of evidence as a contextual variable (Bogacz et al., 2006; Pouget et al., 2013; Shadlen and Kiani, 2013; Glaze et al., 2015), or formulate distinct computational models of 'active' and 'passive' inference (e.g., Gureckis and Markant, 2012; Friston et al., 2013). Our reversal learning task allows us to study the effect of control on inference by comparing two conditions which differ only in the instrumental control over the sampling of evidence. This selective difference changes the nature of uncertainty, but not the complexity of inference across conditions: uncertainty concerns the generative cause of stimuli in the cue-based condition, whereas it concerns the consequence of the previous action in the outcome-based condition. We constructed the task such that the amount of uncertainty is perfectly matched across conditions: Bayes-optimal inference (Glaze et al., 2015) results in the exact same belief trajectories in the two conditions, and thus the exact same behavior. In other words, differences between conditions cannot be attributed to any difference in the evidence provided to participants. Despite this match, we found that interacting with uncertain evidence results in: 1. slower reversal learning, which decreases performance in more volatile environments where reversals are frequent, and 2. higher asymptotic reversal rates, which increases performance in more stable environments where reversals are rare.

While we were primarily interested in differences between conditions, we also found striking similarities in the computations and neural correlates of cue-based and outcome-based inference. At the behavioral level, participants showed correlated inference parameters between conditions. In particular, inference was subject to the same limited precision (Wyart and Koechlin, 2016; Drugowitsch et al., 2016) in the two conditions. These correlations indicate that the two types of inference rely on shared cognitive processes. At the neural level, multiple stimulus properties could be decoded from MEG responses with the same precision in the two conditions, from stimulus orientation in early visual cortex to stimulus evidence in the PPC and LPFC (Heekeren et al., 2008; Hanks and Summerfield, 2017). By applying weights estimated in one condition to decode stimulus properties in the other condition, we further demonstrate that underlying neural 'codes' are shared across the two conditions.

The computational modeling of inference (Glaze et al., 2015) accounted for behavioral differences by an increased reliance on prior beliefs in the outcome-based condition, as if participants perceived a lower rate of reversals than in the cue-based condition. Several alternative accounts can be ruled out based on different aspects of the behavioral and neural data. First, participants did not rely more on prior beliefs in the outcome-based condition because they perceived the incoming evidence as less reliable: they showed equal sensitivity to the presented evidence, and were instructed explicitly that the same probability distributions (categories) were sampled in the two conditions. Furthermore, participants showed no sign of leaky inference within each sequence, in either condition. The slower reversal learning in the outcome-based condition is also unlikely to arise from a biased, choice-supportive filtering of evidence described and reported across cognitive domains (Sharot et al., 2011; Bronfman et al., 2015; Lefebvre et al., 2017; Talluri et al., 2018; Yon et al., 2018). Indeed, the strength of conflicting evidence (inconsistent with beliefs) could be decoded with equal precision across conditions. Finally, the reliance on prior beliefs cannot be explained by a heuristic choice strategy which samples actions either from the prior belief or from the incoming evidence without combining them (Bornstein et al., 2017; Laquitaine and Gardner, 2018). Contrary to such accounts, we could decode from MEG signals the explicit interaction between these two terms: the consistency of incoming evidence with inferred beliefs.

This consistency signal reflects a 'relational' coding of evidence in relation to the current belief about the hidden state being monitored. It is best decoded from MEG signals localized in the MTL, where both spatial and non-spatial cognitive maps have been identified in recent years (Constantinescu et al., 2016; Garvert et al., 2017; Stachenfeld et al., 2017; Vikbladh et al., 2019). Examining the fine temporal dynamics of this neural code revealed that participants update their beliefs following each stimulus, and thus not only when probed for a response. In other words, this coding scheme maps the current stimulus not in relation to the previous decision, but in relation to the current belief. We show that this belief-dependent coding of evidence predicts changes-of-mind in the form of coding switches during the preceding sequence: first stimuli are coded in relation to the belief at the beginning of the trial (before the change-of-mind), whereas later stimuli are coded in relation to the belief at the end of the trial (after the change-of-mind). Importantly, the precision of this neural code scales with the magnitude and direction (sign) of beliefs, thus offering a neural window into their trajectories. The dynamics of this neural code confirmed the lower perceived hazard rate in the outcome-based condition. In particular, coding switches occurred later in the outcome-based condition during sequences leading to changes-of-mind, as predicted by larger prior beliefs which require more conflicting evidence to switch.

In terms of computations, this relational code corresponds to a contextual, history-dependent variable, not directly to the momentary evidence variable being accumulated (Gold and Shadlen, 2007; Hanks et al., 2015; Erlich et al., 2015; Katz et al., 2016). Indeed, its reference point (i.e., the current belief) can switch sign within a single sequence of stimuli, something particularly impractical for downstream read-out. In the outcome-based condition, this means that a non-target stimulus is coded negatively at the beginning of a sequence, but another can be coded positively later in the same sequence if a change-of-mind has occurred between the two stimuli – unlike classical descriptions of reward prediction errors in the midbrain (Schultz et al., 1997; Bayer and Glimcher, 2005). Such dynamic coding scheme is however consistent with recent accounts of opposite coding of factual and counterfactual outcomes in the frontal cortex (Boorman et al., 2009), and of hidden-state inference by dopaminergic signals (Starkweather et al., 2017). Another property of this relational code is that it scales with the magnitude of the current belief: it is almost not expressed when the current belief is weak, but strongly expressed when the current belief is strong – an effect which

violates a basic tenet of Bayesian reasoning: the independence between the likelihood and prior terms (Oaksford and Chater, 2007). Using this relational code as likelihood term (momentary evidence) would result in strong 'circular' inference (Jardri and Denève, 2013; Jardri et al., 2017), a distortion which was not apparent in human behavior.

Interacting with uncertain environments confers several cognitive advantages, which could explain the stabilization of hidden-state inference in the outcome-based condition. First, the active sampling of evidence enables testing current hypotheses in a self-directed, online fashion – an ability known to improve human learning in novel environments (Gureckis and Markant, 2012; Markant and Gureckis, 2014). A functional interpretation is that the lower perceived hazard rate in the outcome-based condition aims at testing the current hypothesis that the previous action samples from the target category. Such hypothesis testing is by definition impossible in the cue-based condition because participants had no control over the sampling of evidence. Note that the inference process used in the cue-based condition is not equivalent to 'observational' learning – i.e., learning from or about another agent interacting with uncertain evidence (Ng and Russell, 2000; Abbeel and Ng, 2004). Indeed, participants were actively committing to interpretations regarding the *source* of evidence in both conditions, the sole difference being the presence or absence of instrumental control over the *sampling* of evidence. Even without considering hypothesis testing, knowledge about action-outcome contingencies affords an agent to predict incoming sensory signals, and to stabilize these signals by stabilizing its behavior (Huys and Dayan, 2009; Haggard and Chambon, 2012; Friston et al., 2013). The lower perceived hazard rate in the outcome-based condition may thus aim at stabilizing the incoming evidence – something which is not possible in the absence of control. Last, the difference in perceived hazard rate between the two conditions may betray a temporal hierarchy between two types of changes which can arise in uncertain environments: 1. external (uncontrollable) changes in the state of the environment (as in the cue-based condition), and 2. changes in the instrumental control of the same environment (as in the outcome-based condition). In this view, participants behave as if they assume changes in control to be less frequent than external changes occurring in equally uncertain environments (Harris, 1996; Klein and Helweg-Larsen, 2002; Wolpe et al., 2014).

Together, these findings bear important consequences for understanding psychiatric dysfunctions of learning and decision-making under uncertainty. In particular, the description of obsessive-compulsive disorder (OCD) as a ‘doubting disease’ fails to distinguish between uncertainty about the cause of external events and uncertainty about the consequence of one’s own actions. Even in the general population, the perceived degree of control over uncertain environments may explain variations in the attitude of individual decision-makers to risk, but also variations in the accuracy of collective decisions whose outcomes shape human societies over long time scales.

Acknowledgments

This work was supported by a starting grant from the European Research Council (ERC-StG-759341) awarded to V.W., a junior researcher grant from the French National Research Agency (ANR-14-CE13-0028-01) awarded to V.W., a France-US collaborative research grant from the French National Research Agency (ANR-17-NEUC-0001-02) and the National Institute of Mental Health (1R01MH115554-01) awarded to V.W. and J.D., and a department-wide grant from the Agence Nationale de la Recherche (ANR-17-EURE-0017). A.W. was supported by the FIRE Doctoral School. V.C. was supported by the French National Research Agency (ANR-16-CE37-0012-01). J.D. was supported by the James S. McDonnell Foundation (grant #220020462).

Author Contributions

Conceptualization: V.W.; Methodology: A.W., V.C., J.D., and V.W.; Formal Analysis: A.W. and V.W.; Investigation: A.W., V.C., and V.W.; Writing – Original Draft: A.W., V.C., and V.W.; Writing – Review & Editing: A.W., V.C., J.D., and V.W.; Supervision: V.W.; Funding Acquisition: J.D. and V.W.

Declaration of Interests

The authors declare no competing interests.

Methods

Lead contact and materials availability

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Valentin Wyart (valentin.wyart@ens.fr).

Experimental model and subject details

24 adult participants took part in the study (12 females, mean age: 24 years, age range: 20–30 years, all right-handed). Participants had no history of neurological and psychiatric disease and had normal or corrected-to-normal vision. All tested participants gave a written informed consent before taking part in the study, which received ethical approval from relevant authorities (Comité de Protection des Personnes Ile-de-France VI, ID RCB: 2007-A01125-48, 2017-A01778-45). Participants received a fixed monetary compensation of 80 euros for their participation in the study.

Method details

Experimental task We asked participants to play a reversal learning task based on visual stimuli which we framed, in two conditions corresponding to different blocks of trials, either as cue-based or outcome-based inference. In both conditions, participants were asked to track a hidden state s of the task which alternates occasionally and unpredictably between two discrete values. Stimuli corresponded to oriented black bars presented in the foreground of a colored disc displaying an angular gradient between orange and blue (through gray) – the two cardinal colors being spaced by $\pi/2$. On each trial, a sequence of 2, 4, 6 or 8 stimuli was drawn from a von Mises probability distribution (category c) centered either on the orientation indicated by orange (A) or the orientation indicated by blue (B) with a fixed concentration κ of 0.5 (Figure 1A). The number of stimuli in each sequence was sampled pseudo-randomly and uniformly across trials. Each sequence was presented at an average rate of 2 Hz, using an inter-stimulus interval of 500 ± 50 ms. The last stimulus of each sequence was followed by a longer delay of 1000 ± 50 ms, before a change in the fixation point probed the participant for a response, by pressing either of two keys with their left or right index finger. All stimuli were presented on a gray background, at a viewing distance of 80 cm. Visual fixation was monitored online

throughout the main task using an EyeLink 1000 eye-tracking system (SR Research, Ottawa, Canada), using a monocular tracking of the dominant eye at a sampling frequency of 1,000 Hz.

Each trial consisted of a stimulus sequence followed by a response (Figure 1B). The task was divided in eight blocks of 72 trials (approx. 7 min. each). At the beginning of each block, an instruction screen framed the upcoming block either as an 'observation' (cue-based) block or as an 'action' (outcome-based) block. The task was presented to participants as a card game, each stimulus depicting a card whose color is determined by the orientation of the black bar relative to the colored background. Each series of cards could be drawn either from the orange deck or from the blue deck, corresponding to the categories A and B. Participants were instructed explicitly that the decks were partially shuffled, such that the orange deck contained mostly orange cards but also blue cards and the blue deck contained mostly blue cards but also orange cards. We also exposed participants to n = 20 draws from each deck (each draw being sampled from a von Mises distribution using the fixed concentration κ of 0.5 used in the main task) before practicing the task.

The description of the 'observation' (cue-based) and 'action' (outcome-based) conditions was fully symmetric, and the order in which participants were instructed about them was counter-balanced across participants. We described the 'observation' (cue-based) condition as follows: "*The computer is drawing cards from one of the two decks. The goal of the game is to identify the deck from which cards are drawn. The computer changes deck from time to time, without warning.*" Each 'observation' block was preceded by an instruction screen indicating the association between keys and decks for the block (left key for the orange deck and right key for the blue deck, or vice versa), counter-balanced across blocks. We described the 'action' (outcome-based) condition as follows: "*You are drawing cards from one of the two decks by pressing a key. The goal of the game is to draw cards only from the target deck, by identifying the key it is associated with. The association between keys and decks reverses from time to time, without warning.*" Each 'action' block was preceded by an instruction screen indicating the target deck for the block (orange or blue), counter-balanced across blocks. Each block was initiated by a response (i.e., a random guess) before the onset of the first stimulus sequence. As can be seen from the descriptions of the two conditions, participants were instructed explicitly about the presence of reversals, but not about their exact

frequency (hazard rate h). Reversals defined 'episodes' during which the hidden-state s (the drawn deck in the 'observation' condition, or the target-drawing key in the 'action' condition) is fixed. The length of these hidden-state episodes was sampled pseudo-randomly from a truncated exponential probability distribution (minimum: 4 trials, maximum: 24 trials) to achieve an approximately constant hazard rate h over the course of each block.

The main task (divided in eight blocks of 72 trials, four of each condition) was preceded by a short practice period (two blocks of 54 trials, one of each condition) in which the concentration κ of stimulus distributions was increased to 1.5 (instead of 0.5 in the main task) such that participants could understand the structure of the task (in particular the presence of reversals in the two conditions) with lower uncertainty regarding the generative category c of each stimulus sequence. Performance in the practice blocks was also used to assess participants' understanding of the two conditions. All tested participants were able to perform both conditions at near-ceiling performance after 54 trials of practice.

Each condition (n = 4 blocks) consisted of two types of blocks: 'more stable' blocks (n = 2) which contained 6 hidden-state episodes (i.e., a hazard rate h of 1/12), and 'more volatile' blocks which contained 12 hidden-state episodes (i.e., a hazard rate h of 1/6). The shorter practice blocks (54 trials instead of 72) contained 6 hidden-state episodes (i.e., a hazard rate h of 1/9, in between the values used in the more stable and more volatile blocks). The eight blocks of the main task were organized in pairs of blocks of the 'observation' and 'action' conditions, one of each volatility, whose order was counter-balanced both within and between participants. The counter-balancing of the different aspects of the task across participants required to test a multiple of n = 8 participants. Given the absence of prior effect sizes for the difference between 'observation' (cue-based) and 'action' (outcome-based) conditions, we chose a sample size (n = 24) which exceeded the average sample size used in human MEG studies at the time of data collection.

Magnetoencephalography MEG data was recorded using a whole-head Elekta Neuromag TRIUX system (Elekta Instrument AB, Stockholm, Sweden) composed of 204 planar gradiometers and 102 magnetometers, at a sampling frequency of 1,000 Hz. Prior to the experiment, each participant's head shape was digitized in the MEG coordinate frame, and four additional head position indicator

(HPI) coils – whose positions were also digitized – were used to monitor and correct for small head movements across blocks. The movement of HPI coils between blocks remained small (mean: 2.3 mm) and did not differ when comparing blocks of the same condition and blocks of different conditions (t_{23} = 1.6, p = 0.128).

As first preprocessing step, magnetic noise from external sources was removed using temporal Signal Space Separation (tSSS), after removing manually detected non-physiological jumps in MEG signals. Stereotyped ocular and cardiac artifacts were corrected using a supervised Principal Component Analysis (PCA) procedure. First, the onset of artifacts (either eye blinks or cardiac R peaks) was detected automatically on auxiliary electrodes (EOG and ECG) synchronized with MEG signals using a threshold-based approach. MEG signals were then epoched from 200 ms before to 800 ms after artifact onsets, and a PCA was used to extract the spatial components of cardiac and ocular artifacts. Typically, one stereotyped PCA component was removed from continuous MEG signals for eye blinks, and two components for heart beats. Continuous MEG signals were high-pass filtered at 0.5 Hz, down-sampled to 500 Hz, and epoched from 200 ms before to 800 ms after the onset of each stimulus (n = 1,440 per participant and per condition). Finally, epoched MEG signals were low-pass filtered at 8 Hz and their analytical representations (decompositions into real and imaginary parts) were computed using the Hilbert transform, resulting in twice the number of MEG signals (408 from planar gradiometers and 204 from magnetometers).

Quantification and statistical analysis

Bayes-optimal inference Let L_t denote the posterior log-odds belief in the current hidden state s_t at trial t, after observing stimulus sequences up to trial t denoted $\Theta_{1:t}$:

$$L_t \equiv \log\left(\frac{p(s_t = 1|\Theta_{1:t})}{p(s_t = 2|\Theta_{1:t})}\right)$$

Given the hazard rate h, as previously derived in (Glaze et al., 2015), the prior log-odds belief at trial $t+1$ corresponds to the following expression:

$$\mathcal{F}(L_t) \equiv \log\left(\frac{1-h}{h} + e^{-L_t}\right) - \log\left(\frac{1-h}{h} + e^{+L_t}\right)$$

Upon observing the new stimulus sequence $\Theta_{t+1} \equiv \{\theta_1, \dots, \theta_n\}$, the belief is updated using Bayes' rule by combining the prior log-odds belief $\mathcal{F}(L_t)$ with the log-odds evidence $\mathcal{L}_{t+1} \equiv \{\ell_1, \dots, \ell_n\}$ provided by the new stimulus sequence, corresponding to:

$$L_{t+1} = \mathcal{F}(L_t) + \mathcal{L}_{t+1}$$

As previously derived in previous work (Drugowitsch et al., 2016), the log-odds evidence $\mathcal{L}_{t+1}$ provided by the new stimulus sequence is defined as:

$$\begin{aligned}
\mathcal{L}_{t+1} &\equiv \sum_{i=1}^{n} \log\left(\frac{p(\theta_i | s_{t+1} = 1)}{p(\theta_i | s_{t+1} = 2)}\right) \\
&= \sum_{i=1}^{n} \kappa \left(\cos\left(2\left(\theta_i - \mu(s_{t+1} = 1)\right)\right) - \cos\left(2\left(\theta_i - \mu(s_{t+1} = 2)\right)\right)\right) \\
&= \sum_{i=1}^{n} 2\,\kappa \cos\left(2\left(\theta_i - \mu(s_{t+1} = 1)\right)\right)
\end{aligned}$$

where $\mu(s_{t+1} = 1)$ indicates the generative mean of the stimulus sequence Θ_{t+1} provided that the current state $s_{t+1} = 1$. In the cue-based condition, $(s_{t+1} = 1) \equiv (c_{t+1} = \mathrm{A})$. Therefore:

$$\mathcal{L}_{t+1} = \sum_{i=1}^{n} 2\,\kappa \cos\left(2(\theta_i - \mu_{\mathrm{A}})\right)$$

In the outcome-based condition, $(s_{t+1} = 1) \equiv \{\mathrm{L} \rightarrow (c_{t+1} = \mathrm{A}), \mathrm{R} \rightarrow (c_{t+1} = \mathrm{B})\}$. Therefore:

$$\mathcal{L}_{t+1} = \begin{cases} \sum_{i=1}^{n} 2\,\kappa \cos\left(2(\theta_i - \mu_{\mathrm{A}})\right) & \text{if } r_t = \mathrm{L}, \\ \sum_{i=1}^{n} 2\,\kappa \cos\left(2(\theta_i - \mu_{\mathrm{B}})\right) & \text{if } r_t = \mathrm{R}. \end{cases}$$

In the cue-based condition, optimal decision-making at trial t corresponds to choosing according to the sign of the posterior log-odds belief L_t. In the outcome-based condition, selecting the action which draws sequences from a target category at trial t corresponds to choosing according to the sign of the prior log-odds belief $\mathcal{F}(L_t)$ carried over at trial $t + 1$. However, by the above expression of $\mathcal{F}(L_t)$, we can guarantee that $\mathcal{F}(L_t) > 0$ as long as $L_t > 0$ and $h < 1/2$, such that it is formally equivalent to cast choices based either on the value of L_t or $\mathcal{F}(L_t)$. Thus, we can use the same Bayes-optimal process in the cue-based and outcome-based conditions.

The fact that the two conditions can be modeled by the same Bayes-optimal process does not guarantee that the two conditions are perfectly matched in terms of provided evidence for any given pair of blocks. Indeed, as can be seen in the above equations, the log-odds evidence $\mathcal{L}_t$ depends on the previous response r_{t-1} in the outcome-based condition, whereas it does not in the cue-based condition. To achieve a perfect match between conditions, we equalized the evidence $\mathcal{L}_{1:\mathrm{T}}$ provided by stimulus sequences, where T is the total number of trials per block ($\mathrm{T} = 72$), using the following procedure. All orientations θ_i appear in the above equations as tilts from a category mean μ. We thus pre-generated each stimulus sequence not as orientations θ_i, but as tilts δ_i from an arbitrary mean. We then defined $\theta_i \equiv \mathrm{mod}(\delta_i + \mu_t, \pi)$, where μ_t is the category mean at trial t. In the cue-based condition, if $s_t = 1$, $\mu_t = \mu_\mathrm{A}$. Therefore:

$$\mathcal{L}_t = \sum_{i=1}^{n} +2\,\kappa\cos(2\,\delta_i)$$

And if $s_t = 2$, $\mu_t = \mu_\mathrm{B}$. Therefore:

$$\mathcal{L}_t = \sum_{i=1}^{n} -2\,\kappa\cos(2\,\delta_i)$$

In the outcome-based condition, if $s_t = 1$,

$$\mu_t = \begin{cases} \mu_\mathrm{A} & \text{if } r_{t-1} = \mathrm{L}, \\ \mu_\mathrm{B} & \text{if } r_{t-1} = \mathrm{R}. \end{cases}$$

Therefore, we obtain the same expression for the log-odds evidence $\mathcal{L}_t$ as in the cue-based condition, now irrespective of the previous response r_{t-1}:

$$\mathcal{L}_t = \sum_{i=1}^{n} +2\,\kappa\cos(2\,\delta_i)$$

And if $s_t = 2$,

$$\mu_t = \begin{cases} \mu_\mathrm{B} & \text{if } r_{t-1} = \mathrm{L}, \\ \mu_\mathrm{A} & \text{if } r_{t-1} = \mathrm{R}. \end{cases}$$

Therefore, we obtain again the same expression as in the cue-based condition:

$$\mathcal{L}_t = \sum_{i=1}^{n} -2\,\kappa\cos(2\,\delta_i)$$

Thus, by using the same pre-generated tilts δ_i for a cue-based block and an outcome-based block, we can perfectly equalize the evidence $\mathcal{L}_{1:\mathrm{T}}$ provided by all stimulus sequences, resulting in the exact

same trajectory of beliefs across conditions. This is because, and despite the dependence of stimulus orientations Θ_t on the previous response r_{t-1} which is constitutive of the outcome-based condition, the evidence $\mathcal{L}_t$ does not depend on r_{t-1} but only on the current hidden state s_t, whose temporal evolution over the course of a block can also be fully matched between conditions.

Noisy Bayesian inference We assume the possible presence of internal variability, or 'noise', at two distinct points of the decision-making process (Drugowitsch et al., 2016). First, inference noise σ_{inf} occurring during the processing of each stimulus sequence – i.e., the updating of beliefs based on the evidence provided by the stimulus sequence (referred to as σ in the main text). And second, selection noise σ_{sel} during response selection – equivalent to a 'softmax' choice policy.

Inference noise reflects additive internal variability in the sequential updating of beliefs based on the evidence provided by each stimulus. Updating the current belief based on n pieces of log-odds evidence introduces n pieces of noise. If each of these pieces of noise is i.i.d. and normally distributed with zero mean and variance σ_{inf}^2, then the noise introduced by the processing of the sequence has variance $n\,\sigma_{inf}^2$. As described in previous work (Drugowitsch et al., 2016), inference noise turns the deterministic belief update equation into a stochastic draw:

$$L_t \sim \mathcal{N}\left(\mathcal{F}(L_{t-1}) + \mathcal{L}_t, n\,\sigma_{\text{inf}}^2\right).$$

Selection noise does not perturb the inference process itself, and only corrupts response selection. We model such internal variability by sampling responses from the sign of a normally distributed decision variable with mean L_t and variance σ_{sel}^2. In some analyses, we also considered the presence of response lapses – i.e., trials in which the participant blindly repeats the previous response instead of selecting the response based on the posterior belief. This leads to a fraction p_{lapse} of trials associated with a probability of repeating the previous response $p(r_t = r_{t-1}) = 1$.

Model-based analysis of behavior For a block of T trials, each fully specified by its stimulus sequence Θ_t, we observed a series of responses $r_1, \ldots, r_{\mathrm{T}}$. We fitted the parameters values $\phi \equiv \{h, \sigma_{\text{inf}}, \sigma_{\text{sel}}, p_{\text{lapse}}\}$ that resulted in the best match between the observed series of responses and the series of responses predicted by the model, $p(r_{1:\mathrm{T}}|\Theta_{1:\mathrm{T}}, \phi)$. We also fitted 'reduced' models that

removed certain components by fixing their corresponding parameters to zero. For example, a model without selection noise would correspond to $\sigma_{\text{sel}} = 0$.

We found best-fitting parameter values by sampling from the Bayesian posterior over parameters using particle Monte Carlo Markov Chain (MCMC) methods (Andrieu et al., 2010). These methods use standard MCMC methods to sample from the parameter posterior, but replace computation of the parameter likelihood $p(r_{1:\text{T}}|\Theta_{1:\text{T}}, \phi)$ – not possible in closed form – with a noisy but unbiased approximation of this likelihood by a particle filter. As MCMC method we used the adaptive mixture Metropolis method (Roberts and Rosenthal, 2009) that adapts its proposal distribution in an initial burn-in period to achieve favorable acceptance ratios. Prior distributions over parameters $p(\phi)$ were defined as truncated normal distributions (h: mean = 0.2, s.d. = 0.1, range = [0,1]; σ_{inf}: mean = 0.5, s.d. = 0.2, range = [0,10]; σ_{sel}: mean = 0.5, s.d. = 0.2, range = [0,10]; p_{lapse}: mean = 0.01, s.d. = 0.05, range = [0,1]). Note that the particle filter also provided an unbiased estimate of the belief trajectory $p(L_{1:t}|r_{1:t}, \Theta_{1:t}, \phi)$ for $t \in [1,\text{T}]$, which was used in subsequent analyses of MEG signals. The particle MCMC fitting procedure was written in Julia v1.0 (Bezanson et al., 2017; https://julialang.org), and run using 10,000 particles.

We used the unbiased estimate of the marginal likelihood $p(r_{1:\text{T}}|\Theta_{1:\text{T}})$ computed using conditional likelihoods from the particle filter (Andrieu et al., 2010) as model evidence in Bayesian Model Selection (BMS) analyses. This metric integrates over parameters ϕ and thus penalizes model complexity (Pitt and Myung, 2002) without requiring explicit penalization terms as in the Bayesian Information Criterion (BIC). BMS was conducted using separate fixed-effects and random-effects approaches – which yielded qualitatively identical results. The fixed-effects approach assumes that all participants are relying on the same model, and consists in comparing the log-marginal likelihood summed across participants for each tested model. By contrast, the random-effects approach assumes that different participants may rely on different models, and consists in estimating the distribution over models that participants draw from (Stephan et al., 2009). We used the Dirichlet parameterization of the random-effects approach implemented in SPM12 (Wellcome Center for Human Neuroimaging, London, UK; http://www.fil.ion.ucl.ac.uk/spm).

Psychometric analysis of behavior Reversal learning behavior was characterized by two psychometric curves. First, we fitted response *reversal* curves – i.e., the probability of correctly identifying the hidden state of the current episode as a function of trial position in the current episode – by a saturating exponential function:

$$p(r_k = s) = p_0 + (p_{\text{rev}} - p_0)\left(1 - e^{-\frac{k}{t_{\text{rev}}}}\right)$$

where $p(r_k = s)$ corresponds to the probability of correctly identifying the current hidden state s at trial k, with k the position of the trial in the current episode. The two fitted parameters are the reversal time constant t_{rev} and the asymptotic reversal rate p_{rev}. The initial reversal rate p_0 is set to 0.5 for the first episode of each block, and to $1 - p(r_k = s)$ reached at the last trial of the previous episode for subsequent episodes. Best-fitting parameter values for t_{rev} and p_{rev} were obtained by maximum likelihood estimation (MLE) through gradient descent (on the negative log-likelihood) using the interior point algorithm implemented in MATLAB (MathWorks, Natick, MA, USA).

Second, we fitted response *repetition* curves – i.e., the probability of repeating the previous response as a function of the evidence provided by the intervening sequence in favor of the previous response – by a three-parameter sigmoid function:

$$p(r_t = r_{t-1}) = p_{\text{rep}} + \left(1 - p_{\text{rep}}\right)\frac{1}{1 + e^{-\beta\mathcal{L}_{\text{rep},t} + \beta_{\text{rep}}}}$$

where $p(r_t = r_{t-1})$ corresponds to the probability of repeating the previous response and $\mathcal{L}_{\text{rep},t}$ corresponds to the log-odds evidence provided by the stimulus sequence at trial t in favor of the previous response r_{t-1}:

$$\mathcal{L}_{\text{rep},t} \equiv \sum_{i=1}^{n} \log\left(\frac{p(\theta_i | s_t = r_{t-1})}{p(\theta_i | s_t \neq r_{t-1})}\right)$$

The three fitted parameters are the slope of the sigmoid curve β (indexing the sensitivity of responses to the evidence $\mathcal{L}_{\text{rep},t}$), the decision criterion for repeating the previous response β_{rep}, and the lower asymptote of the sigmoid curve p_{rep} (reflecting the fraction of evidence-independent repetitions). As above, best-fitting parameter values were obtained by maximum likelihood estimation (MLE). The decision criterion β_{rep} was reparametrized as the point of subjective equivalence (PSE) – i.e., the amount of evidence $\mathcal{L}_{\text{rep},t}$ for which $p(r_t = r_{t-1})$ = 0.5.

Multivariate pattern analyses of MEG signals We applied multivariate pattern analyses to stimulus-locked MEG epochs to estimate the neural patterns associated with four characteristics of each stimulus i, where i corresponds to the position of the stimulus in question in the current sequence: 1. its orientation θ_i described by $\cos(2\,\theta_i)$ and $\sin(2\,\theta_i)$, 2. its change (tilt) from the previous stimulus, described by $|\theta_i - \theta_{i-1}|$, 3. the strength of the evidence provided by the stimulus, described by its tilt from the nearest category boundary θ_0, $|\theta_i - \theta_0|$, and 4. the consistency of the stimulus with the previous response r_{t-1}, described as its tilt from the nearest category boundary signed by the direction of the provided evidence ℓ_i in relation to r_{t-1}. In some analyses, we also defined consistency with respect to the current belief $L_{t,i}$ – i.e., accounting for previous stimuli in the current sequence, in which case the tilt from the nearest category boundary was signed by the direction of the provided evidence ℓ_i in relation to $L_{t,i}$. Importantly, owing to features of our experimental task (in particular the changes in category means μ_{A} and μ_{B} across trials), these four stimulus characteristics showed no significant correlation with each other for a given stimulus i nor between successive stimuli i and $i+1$ (all *r*-squared < 0.001).

We used a cross-validated multivariate linear encoding model to estimate the spatial MEG patterns $\widehat{w}$ associated for each participant with each stimulus characteristic x at every time point from 100 ms before to 700 ms after stimulus onset. For each cross-validation fold (n = 12, interleaved), we defined spatial MEG patterns $\widehat{w}$ on the training set by regressing – in a least-squares sense – each z-scored gradiometer feature z_{train} (real and imaginary parts, yielding 408 features in total) against the stimulus characteristic x_{train} across stimulus exemplars (n = 1,320 for each cross-validation fold), by solving $\mathrm{argmin}_w(\|z_{\mathrm{train}} - w\,x_{\mathrm{train}}\|^2)$ for each MEG feature. We then projected the MEG data on the test set $\mathcal{Z}_{\mathrm{test}}$ on the dimension defined by the coding weights $\widehat{w}$ to obtain neural predictions $\hat{x}_{\mathrm{test}}$ of the stimulus characteristic x_{test} for each epoch of the test set (n = 120 for each cross-validation fold). After applying this procedure for each cross-validation fold, we computed the linear correlation coefficient between neural predictions $\hat{x}$ and ground-truth values x of the stimulus characteristic. The coding precision metric reported in the main text corresponds to the Fisher transform of the correlation coefficient, which is approximately normally distributed, such that we could compute standard parametric statistics at the group level.

The multivariate pattern analyses described above were conducted for each participant and each condition. At the group level, we used standard parametric tests (paired *t*-tests, repeated-measures ANOVAs) to assess the statistical significance of observed differences in coding precision between conditions across tested participants. Neural coding latency was computed for each stimulus characteristic and each condition by estimating the peak of coding precision using a jackknifing (leave-one-out) procedure (Kiesel et al., 2008). The type 1 error rate arising from multiple comparisons was controlled for using non-parametric cluster-level statistics computed across time points (Maris and Oostenveld, 2007). All findings reported in the main text were robust to changes in the method used for computing spatial MEG patterns (e.g., by applying ridge-regression decoding instead of least-squares encoding) and to the number of cross-validation folds.

Source reconstruction of MEG signals Cortical surface-based segmentation and reconstruction was performed with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu) using the default surface-based pipeline (Dale et al., 1999; Fischl et al., 1999), based on high-resolution T1-weighted anatomical MRIs recorded from each participant using a Siemens Verio 3T scanner (Siemens AG, Munich, Germany) and a 64-channel head coil, driven by a 3D MPRAGE sequence with an isotropic resolution of 1 mm^3. Obtained anatomical surfaces were down-sampled to 5,000 vertices and realigned to the MEG spatial coordinate frame using each participant's digitized head shape and fiducial markers. Source reconstruction was performed with Brainstorm (Tadel et al., 2011; http://neuroimage.usc.edu/brainstorm) using weighted minimum norm estimation (MNE) with Tikhonov regularization (Baillet et al., 2001). Forward modeling was performed using overlapping spheres. We obtained noise covariance matrices from three minutes of 'empty room' recordings (Gross et al., 2013) on which we applied the exact same preprocessing pipeline, using a diagonal regularization λ of 0.1. Vertex-wise inverse operators were computed based on MEG signals from both gradiometers and magnetometers (non-Hilbert transformed) by constraining dipole orientations to be orthogonal to the cortical surface.

Source reconstruction was conducted for each participant based on his/her individual cortical surface. The results of analyses performed in source space were then co-registered to a common anatomy (the ICM152 MNI template) for group-level statistics based on the FreeSurfer registration

procedure. In practice, vertex-level results obtained for each participant were interpolated from the subject-specific anatomy to the common anatomy using inverse distance weighting as implemented in Brainstorm (Shepard's method).

Searchlight analysis of MEG signals We conducted 'searchlight'-based multivariate pattern analyses on the activity of reconstructed cortical sources (vertices) from each participant (Etzel et al., 2013). After decomposing the activity of each vertex into its real and imaginary parts using the Hilbert transform (as done for sensor-level analyses), we computed coding precision at each vertex position and each time point around stimulus onset by using as features the current vertex and its ten closest neighbors in terms of Euclidean distance. To increase the spatial selectivity of searchlight-based analyses, we computed for each stimulus characteristic and each vertex position a selectivity metric, corresponding to the result of a jackknifed *F*-test which compares coding precision of this characteristic to the coding precision of other characteristics. We accounted for overall differences in coding precision between characteristics by normalizing the coding precision of each characteristic at each vertex position by the peak coding precision of this characteristic across all time points and vertices. Non-parametric cluster-level statistics (Maris and Oostenveld, 2007) for searchlight-based analyses were computed across time points and vertices as implemented in FieldTrip (Oostenveld et al., 2011; http://www.fieldtriptoolbox.org). We marginalized these cluster-level statistics across either space or time by summing statistics across the corresponding dimension for display purposes.

Cross-condition generalization analysis of MEG signals We conducted cross-condition generalization analyses of spatial MEG patterns to assess the degree of similarity between neural representations of stimulus characteristics across conditions. In practice, we used the coding weights estimated on the training set of one condition to compute neural predictions on the test set of the other condition. Although these cross-condition generalization analyses do not require cross-validation to avoid overfitting since the two conditions are non-overlapping, we wanted to compare coding precision scores between within-condition (non-generalized) and between-condition (generalized) settings. We thus applied the same 12-fold cross-validation approach within each condition (which had the same number of stimulus exemplars) such that: 1. coding weights $\hat{w}$ were estimated using

the same amount of training data in within- and between-condition settings, and 2. coding precision was computed using the same amount of test data in the two settings.

To quantify the degree of similarity between two neural codes based on the results of cross-condition generalization analyses, we simulated and fitted a fixed number n of linear coding units z_i (n = 300) to within- and between-condition coding precision scores obtained from the MEG data. A fixed number n_{sel} of these units (n_{sel} = 100) was defined as 'selective' to a stimulus characteristic x, meaning that their activity scales linearly with x (Figure 5A):

$$z_i = \beta_i\, x + \varepsilon$$

where β_i = 1 for selective units and 0 for non-selective units, and $\varepsilon \sim \mathcal{N}(0, \sigma)$ corresponds to background noise of variance σ^2 (i.i.d. across units). We used the same cross-validated multivariate linear encoding model, and the same amount of data, to estimate coding patterns $\hat{w}$ for the stimulus characteristic x from population activity and obtain predictions $\hat{x} \equiv \hat{w}^{\mathrm{T}}\, Z$. Simulating the same population of units in two conditions was controlled by three parameters: the amount σ of background noise in each condition (σ_{Cb} and σ_{Ob}), and the fraction of units ω selective to the stimulus characteristic x in both conditions (coding similarity). Parameters σ_{Cb} and σ_{Ob} scaled negatively with coding precision within each condition (values 1-2 in Figure 5), whereas parameter ω scaled positively with coding precision between conditions (values 3-4 in Figure 5). We fitted all three parameters to the four coding precision values obtained from the cross-condition generalization analysis using a jackknifed least-squared error minimization procedure. Importantly, the best-fitting estimate of coding similarity ω was robust to changes in the total number n of units, and to the number n_{sel} of selective units (both fixed across conditions).

Estimation of neural hazard rate from MEG signals We used the estimate of the belief trajectory $p(L_{1:t}|r_{1:t}, \Theta_{1:t}, \phi)$ for $t \in [1,\mathrm{T}]$ provided by the particle filter – conditioned on each participant's responses $r_{1:t}$, stimulus sequences $\Theta_{1:t}$ and model parameters ϕ to fit the 'neural' hazard rate h^* for which we could best decode consistency from MEG signals. For this purpose, we computed belief trajectories for varying values of the perceived hazard rate h (Figure 7A) – i.e., ignoring the best-fitting value obtained from behavior – and derived for each of these trajectories the consistency of each stimulus in relation to the estimated belief conditioned on each value of h. These alternative

instantiations of the same consistency variable x_{cons} were used to compute 'tuning curves' of coding precision as a function of h. We then fitted the hazard rate value $h^* \equiv \text{argmax}_h\left(\text{corr}(\hat{x}_{\text{cons}}, x_{\text{cons}})\right)$ for each participant and condition, where x_{cons} depends implicitly on h and $\hat{x}_{\text{cons}}$ corresponds to neural predictions obtained using the multivariate pattern analysis described above. To obtain smooth estimates of h^*, we approximated each tuning curve by a quadratic polynomial function using a least-squared error minimization procedure and computed its maximum. Importantly, using unsmoothed estimates of h^* did not change the pattern of findings.

Data and code availability

The manuscript includes all datasets generated or analyzed during this study. The data and analysis code supporting the reported findings will be available from the Lead Contact upon request upon publication of the manuscript in a peer-reviewed journal.

References

Abbeel, P., and Ng, A.Y. (2004). Apprenticeship learning via inverse reinforcement learning. In *Proc. ICML 21* (Banff, Alberta, Canada).

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. B* **72**, 269–342.

Baillet, S., Mosher, J.C., and Leahy, R.M. (2001). Electromagnetic brain mapping. *IEEE Signal Process. Mag.* **18**, 14–30.

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

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47**, 129–141.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. (2017). Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**, 65–98.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J.D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765.

Boorman, E.D., Behrens, T.E.J., Woolrich, M.W., and Rushworth, M.F.S. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron* **62**, 733–743.

Bornstein, A.M., Khaw, M.W., Shohamy, D., and Daw, N.D. (2017). Reminders of past choices bias decisions for reward in humans. *Nat. Commun.* **8**, 15958.

Bronfman, Z.Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., and Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proc. R. Soc. B* **282**, 20150228.

Constantinescu, A.O., O'Reilly, J.X., and Behrens, T.E.J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468.

Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical surface-based analysis. I: segmentation and surface reconstruction. *Neuroimage* **9**, 179–194.

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., and Koechlin, E. (2016). Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1398–1411.

Erlich, J.C., Brunton, B.W., Duan, C.A., Hanks, T.D., and Brody, C.D. (2015). Distinct effects of prefrontal and parietal cortex inactivations on an accumulation of evidence task in the rat. *eLife* **4**, 05457.

Etzel, J.A., Zacks, J.M., and Braver, T.S. (2013). Searchlight analysis: promise, pitfalls, and potential. *Neuroimage* **78**, 261–269.

Fischl, B., Sereno, M.I., and Dale, A.M. (1999). Cortical surface-based analysis. II: inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9**, 195–207.

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., and Dolan, R.J. (2013). The anatomy of choice: active inference and agency. *Front. Hum. Neurosci.* **7**, 598.

Garvert, M.M., Dolan, R.J., and Behrens, T.E.J. (2017). A map of abstract relational knowledge in the human hippocampal--entorhinal cortex. *eLife* **6**, 17086.

Glaze, C.M., Kable, J.W., and Gold, J.I. (2015). Normative evidence accumulation in unpredictable environments. *eLife* **4**, 08825.

Gold, J.I., and Shadlen, M.N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci.* **30**, 535–574.

Grootswagers, T., Wardle, S.G., and Carlson, T.A. (2017). Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J. Cogn. Neurosci.* **29**, 677–697.

Gross, J., Baillet, S., Barnes, G.R., Henson, R.N., Hillebrand, A., Jensen, O., Jerbi, K., Litvak, V., Maess, B., Oostenveld, R., et al. (2013). Good practice for conducting and reporting MEG research. *Neuroimage* **65**, 349–363.

Gureckis, T.M., and Markant, D.B. (2012). Self-directed learning: a cognitive and computational perspective. *Perspect. Psychol. Sci.* **7**, 464–481.

Haggard, P., and Chambon, V. (2012). Sense of agency. *Curr. Biol.* **22**, R390–R392.

Hanks, T.D., Kopec, C.D., Brunton, B.W., Duan, C.A., Erlich, J.C., and Brody, C.D. (2015). Distinct relationships of parietal and prefrontal cortices to evidence accumulation. *Nature* **520**, 220–223.

Hanks, T.D., and Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. *Neuron* **93**, 15–31.

Harris, P. (1996). Sufficient grounds for optimism? The relationship between perceived controllability and optimistic bias. *J. Soc. Clin. Psychol.* **15**, 9–52.

Heekeren, H.R., Marrett, S., and Ungerleider, L.G. (2008). The neural systems that mediate human perceptual decision making. *Nat. Rev. Neurosci.* **9**, 467–479.

Huys, Q.J.M., and Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition* **113**, 314–328.

Izquierdo, A., Brigman, J.L., Radke, A.K., Rudebeck, P.H., and Holmes, A. (2017). The neural basis of reversal learning: an updated perspective. *Neuroscience* **345**, 12–26.

Jardri, R., and Denève, S. (2013). Circular inferences in schizophrenia. *Brain* **136**, 3227–3241.

Jardri, R., Duverne, S., Litvinova, A.S., and Denève, S. (2017). Experimental evidence for circular inference in schizophrenia. *Nat. Commun.* **8**, 14218.

Katz, L.N., Yates, J.L., Pillow, J.W., and Huk, A.C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. *Nature* **535**, 285–288.

Kiesel, A., Miller, J., Jolicœur, P., and Brisson, B. (2008). Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods. *Psychophysiology* **45**, 250–274.

Klein, C.T.F., and Helweg-Larsen, M. (2002). Perceived control and the optimistic bias: a meta-analytic review. *Psychol. Health* **17**, 437–446.

Laquitaine, S., and Gardner, J.L. (2018). A switching observer for human perceptual estimation. *Neuron* **97**, 462–474.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., and Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nat. Hum. Behav.* **1**, 0067.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190.

Markant, D.B., and Gureckis, T.M. (2014). Is it better to select or to receive? Learning via active and passive hypothesis testing. *J. Exp. Psychol. Gen.* **143**, 94–122.

Ng, A.Y., Russell, S.J., and Others (2000). Algorithms for inverse reinforcement learning. In *Proc. ICML 17* (Stanford, CA, USA).

Oaksford, M., and Chater, N. (2007). *Bayesian rationality: the probabilistic approach to human reasoning* (Oxford University Press).

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 156869.

Pitt, M.A., and Myung, I.J. (2002). When a good fit can be bad. *Trends Cogn. Sci.* **6**, 421–425.

Pouget, A., Beck, J.M., Ma, W.J., and Latham, P.E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* **9**, 545–556.

Ratcliff, R., and Smith, P.L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev.* **111**, 333–367.

Roberts, G.O., and Rosenthal, J.S. (2009). Examples of adaptive MCMC. *J. Comput. Graph. Stat.* **18**, 349–367.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron* **70**, 1054–1069.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. *Science* **275**, 1593–1599.

Shadlen, M.N., and Kiani, R. (2013). Decision making as a window on cognition. *Neuron* **80**, 791–806.

Shadlen, M.N., and Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron* **90**, 927–939.

Sharot, T., Korn, C.W., and Dolan, R.J. (2011). How unrealistic optimism is maintained in the face of reality. *Nat. Neurosci.* **14**, 1475–1479.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017.

Sutton, R.S., and Barto, A.G. (1998). *Reinforcement learning: an introduction* (MIT Press).

Sutton, R.S., and Barto, A.G. (1998). *Reinforcement learning: an introduction – second edition* (MIT Press).

Stachenfeld, K.L., Botvinick, M.M., and Gershman, S.J. (2017). The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653.

Starkweather, C.K., Babayan, B.M., Uchida, N., and Gershman, S.J. (2017). Dopamine reward prediction errors reflect hidden-state inference across time. *Nat. Neurosci.* **20**, 581–589.

Tadel, F., Baillet, S., Mosher, J.C., Pantazis, D., and Leahy, R.M. (2011). Brainstorm: a user-friendly application for MEG/EEG analysis. *Comput. Intell. Neurosci.* **2011**, 879716.

Talluri, B.C., Urai, A.E., Tsetsos, K., Usher, M., and Donner, T.H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Curr. Biol.* **28**, 3128–3135.

Vikbladh, O.M., Meager, M.R., King, J., Blackmon, K., Devinsky, O., Shohamy, D., Burgess, N., and Daw, N.D. (2019). Hippocampal contributions to model-based planning and spatial memory. *Neuron* **102**, 683–693.

Wolpe, N., Wolpert, D.M., and Rowe, J.B. (2014). Seeing what you want to see: priors for one's own actions represent exaggerated expectations of success. *Front. Behav. Neurosci.* **8**, 232.

Wyart, V., and Koechlin, E. (2016). Choice variability and suboptimality in uncertain environments. *Curr. Opin. Behav. Sci.* **11**, 109–115.

Yon, D., Gilbert, S.J., de Lange, F.P., and Press, C. (2018). Action sharpens sensory representations of expected outcomes. *Nat. Commun.* **9**, 4288.