Distinguishing neural correlates of prediction errors on perceptual content and detection of content

Abbreviated title: Content and detection prediction errors

Nadine Dijkstra[*1], Oliver Warrington[1], Peter Kok[1] and Stephen M. Fleming[1,2,3]

1. Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, University College London, United Kingdom, WC1N 3AR
2. Max Planck UCL Centre for Computational Psychiatry and Aging Research, University College London, United Kingdom, WC1B 5EH
3. Department of Experimental Psychology, University College London, United Kingdom, WC1H 0AP

* Corresponding author

Contact: n.dijkstra@ucl.ac.uk

Conflict of interest statement: The authors declare to have no conflict of interest.

Accounting for why discrimination between different perceptual contents is not always accompanied conscious detection of that content remains a challenge for predictive processing theories of perception. Here we test a hypothesis that detection is supported by a distinct inference within generative models of perceptual content. We develop a novel visual perception paradigm that probes such inferences by manipulating both expectations about stimulus content (stimulus identity) and detection of content (stimulus presence). In line with model simulations we show that both content and detection expectations influence reaction times on a categorisation task. By combining a no-report version of our task with functional neuroimaging we reveal that violations of expectations (prediction errors; PEs) about perceptual content and detection are supported by posterior and prefrontal cortex in qualitatively different ways: within posterior sensory cortex, activity patterns diverge only on trials with a content PE, but within these trials, further divergence is seen for detection PEs. In contrast, within prefrontal cortex, activity patterns diverge only on trials with a detection PE, but within these trials, further divergence is seen for content PEs. These results suggest rich encoding of both content and detection prediction errors and highlight a distributed neural basis for inference on content and detection of content in the human brain.

Our perceptual experience is characteristically limited: at any given moment in time we are aware of only a subset of perceptual inputs. Such failures of awareness do not necessarily reflect failures of sensory processing. For instance, in a widely cited example of a dissociation between perceptual performance and awareness, a patient with blindsight is still able to respond above-chance to the identity of a stimulus, despite not seeing that stimulus (Persaud et al., 2011; Weiskrantz et al., 1974). Similar dissociations have been documented in otherwise healthy subjects using techniques such as masking, where the content of stimuli which are rendered invisible nevertheless continues to exert an impact on behaviour (Dehaene et al., 2001; Marcel, 1983; Peters & Lau, 2015, although see (Meyen et al., 2022). Within the framework of perceptual decision-making, dissociations between performance and awareness can be modelled as a distinction between discrimination – categorising some aspect of stimulus identity – and detection – responding as to whether a stimulus is perceived or not (Azzopardi & Cowey, 1997; Green & Swets, 1966; Peters & Lau, 2015).

Predictive processing offers a powerful and general computational framework for modelling perception (Hohwy & Seth, 2020; Marvan & Havlík, 2021). Within this framework, the content of perception is realized by combining prior knowledge (expectations) with incoming sensory evidence (Bastos et al., 2012; Friston et al., 2006; Kersten et al., 2004; Kok et al., 2013). Mismatches between expectations and evidence about a particular feature result in prediction errors – tell-tale signatures of inference on that feature. Previous accounts have suggested that what we become aware of is determined by specific aspects of perceptual inference, for example, the perceptual hypothesis with the highest posterior probability (Hohwy, 2012) or the updating of perceptual hypotheses by unexpected signals (Hobson & Friston, 2014). However, these accounts struggle to accommodate

dissociations between high-fidelity discrimination performance – presumably reflecting intact perceptual inference – and detection judgments (Lau, 2022).

An alternative proposal is that detection arises from inferences that are distinct from inferences about content. We recently proposed a computational architecture (the higher-order state space (HOSS) model) in which a higher-order global inference about the presence or absence of first-order perceptual content supports detection judgements (Lau, 2019; Lau, 2007; Morales, 2022). This processing step is proposed to be distinct to bottom-up salience or attention (Fleming, 2020). This model builds on a large body of prior work that associates awareness with changes in higher-order cognitive processes, including global workspace and higher-order theories of consciousness (Brown, 2015; Dehaene & Changeux, 2011; Lau & Rosenthal, 2011; Mashour et al., 2020). A more general question that goes beyond these theories of consciousness is whether inferences on content and detection of content rely on distinct processes in the human brain. Addressing this question would provide initial empirical constraints on the architecture of predictive processing theories of perception and consciousness.

In this study, we investigated to what extent inferences about content and detection of content correlate with distinct neural substrates measured with fMRI. In what follows, we refer to predictions about perceptual content that are relevant for discrimination as "content expectations", and to predictions about the presence (vs. absence) of content that are relevant for detection as "detection expectations". We developed a novel experimental paradigm in which we independently manipulated expectations about perceptual content (stimulus identity) and detection (whether stimulus content will be present or absent). In line with a neural hierarchy supporting detection inferences, we hypothesised that prediction errors about perceptual content would be localized to sensory cortex (Bastos et al., 2012; Kok et al., 2013) whereas prediction errors on detection would be localized to prefrontal cortex (Merten & Nieder, 2012; van Vugt et al., 2018).

To preface our results, we first show in a behavioural experiment that both content and detection expectations influence reaction times. Using a no-report version of the same task in conjunction with neuroimaging (Tsuchiya et al., 2015), we show that content prediction errors are predominantly encoded in sensory (visual) cortical areas whereas detection prediction errors are predominantly encoded in prefrontal cortical areas. However, a strict separation between content and detection of content is nuanced by findings of mutual interactions between the two types of prediction error signals in both visual and prefrontal cortices. Taken together, our findings suggest that inferences on content and detection of content rely on distinct but interacting neural substrates in the human brain.

Materials and Methods

Participants. To determine our sample size, we assumed medium effect sizes (Cohen's d = 0.05) which require 34 participants to achieve a power of 80% to detect an effect at an alpha level of 0.05. To allow for drop-out, 36 participants gave written informed consent and participated in the study (mean age 26.4, SD 7.2). All participants were included in the behavioural analysis. 9 participants were excluded from the MRI analyses: 2 participants took part in the behavioural session but could not be scanned, 6 were excluded due to low catch trial accuracy (below 70%), 1 was excluded due to a technical error during response recording. 27 participants were included in the final fMRI analyses. Given that dropout for the MRI session was higher than anticipated, this reduced our power to detect a medium effect size to 71%. The study was approved by the University College London ethics committee (approval number 8231_001). Participants were paid £8 for the behaviour session and £10 per hour for the fMRI session.

Stimuli. The experiment was programmed in MATLAB R2019 (MathWorks) using Psychtoolbox (version 3.0.16). In the behavioural session, stimuli were presented on a desktop monitor and in the scanner, stimuli were presented via a projector at an approximate viewing distance of 58cm. Stimuli consisted of 8 faces and 8 houses displayed in black and white and embedded in noise. The face stimuli were selected from the Karolinska Directed Emotional Faces database (Lundqvist, D., Flykt, A., & Öhman, 1998) with a 50-50 male-female ratio and a neutral expression. The house stimuli were adapted from the Pasadena houses database collected by Helle and Perona (California Institute of Technology, Pasadena, California). We controlled for differences in spatial frequency between stimuli by calculating the frequency phase and magnitude for each stimulus in the set using Fourier transformation, and then combining the mean spatial frequency magnitude over stimuli with each individual image's frequency phase information to create images that are perceptually similar to the original images but have the same spatial frequency profile. Noise was generated as matrices of random numbers uniformly distributed between 0 and 1 and combined with each stimulus image in a weighted sum with the weight on the stimulus image controlled by a visibility parameter. Face and house stimuli were each presented at 0.9 visibility. Noise images consisted solely of noise. To ensure participant engagement in the fMRI session, 20% of the trials were catch trials in which 10% of image pixels were turned green. Participants were instructed to press a button when they saw a green tinge to the image (more details below).

Task and procedure. Participants attended the laboratory on two different days within a 5-day window. On the first day they completed a behavioural training session in which the task and cue-

stimulus associations were learned (Fig. 1A). On the second day participants completed the fMRI session during which they performed the same task but now without requiring a response (Fig. 1B). To facilitate investigation of whether prediction errors on detection are encoded differently from prediction errors on content, we independently manipulated the probability of whether any perceptual content (face or house) would be presented or not (the detection prior), and whether such content would be a face or a house (the content prior). These probabilities were reflected in the shape (detection) and colour (content) of the cues (Fig. 1C). For example, an orange circle indicates that most likely no stimulus would be presented, but that if one were presented, it would likely be a house. To prevent neural correlates of predictions being contaminated by responses to the physical cues, cue-stimulus mappings within each level were swapped halfway through the experiment. For instance, the high-probability presence / high-probability face cue could be represented by a blue rectangle in the first half of the experiment, and an orange circle in the second half. Both mappings were also used during the behavioural training session to familiarize participants with possibility of such switches.

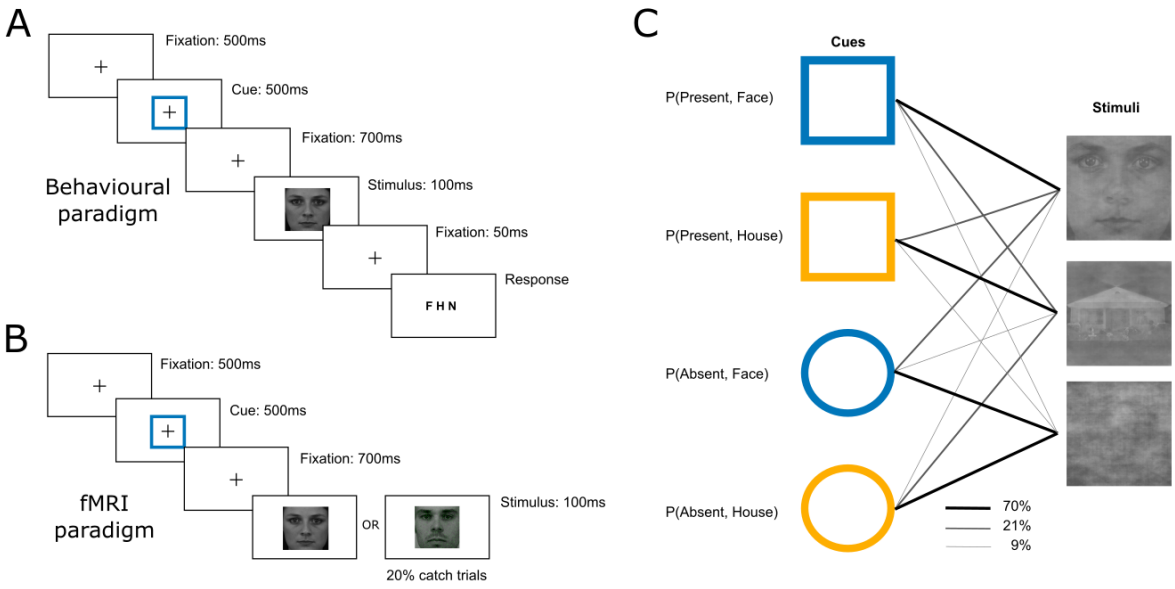

Figure 1. Experimental paradigm. (**A**) Behavioural paradigm. Trials consisted of 500ms fixation followed by 500ms cue, 700ms fixation, 100ms stimulus. After another 50ms fixation participants had to indicate whether they saw a face (F), house (H) or noise (N) using the 'a', 's' or 'd' keyboard keys respectively. (**B**) The fMRI paradigm was the same as the behavioural paradigm except that no response was required, ensuring task-related activations would not reflect reporting requirements. To ensure participants remained engaged with the task, they were instructed to press a button when the stimulus contained green pixels, which was the case on 20% of the trials. (**C**) The shape of the cue indicated the probability a stimulus would be present or not (detection expectation) whereas the colour indicated whether that stimulus was likely to be a face or a house (content expectation). For example, a blue circle indicated a high probability that no stimulus would be presented but that if a stimulus were presented, it would likely be a face.

Behavioural session. At the start of the experiment, participants were trained on the task and the relationship between the cues and stimuli in 3 phases. Within each phase, task instructions were

presented to participants via a self-paced PowerPoint presentation interleaved with 20 practice trials, resulting in 60 practice trials in total. In phase 1, participants were introduced to how the colour of cues indicated stimulus content (e.g. blue: '70% chance of being a face', orange: '70% change of being a house'). During each trial, a cue was presented for 500ms, followed by a fixation cross for 700ms, and then the stimulus for 100ms (Fig. 1A). After another brief 50ms fixation, a response screen appeared indicating the different response options (in this phase, only 'F' for face and 'H' for house responses were available). Participants were instructed to use the 'a' and 's' keys to indicate whether they saw a face or a house respectively. After pressing the button, their selected category, 'F' or 'H', was highlighted for 200ms before continuing to the next trial.

In phase 2, the noise category (absence of content) was added and participants were introduced to the shape dimension of the cues (e.g. square: '70% likely to be a picture (face or house)', rectangle: '70% likely to be no picture/noise'). They then practiced 20 trials of face-house-noise categorisation with just the presence/absence shape cues. 'N' was added to the response options and participants used the 'd' key to select this 'no picture' answer. In the final phase 3, both cue dimensions were combined and their meaning was explained (e.g. blue square: '70% likely to be a picture AND that picture will likely be a *face*'). Then, participants completed two blocks of the main task with these cues, each taking approximately 6 minutes. After this, the cue-stimulus mappings were swapped, the new cues were explained and another similar training session was completed. Finally, the participants completed two blocks with these new cue-stimulus mappings.

fMRI session. To ensure that any correlates of prediction errors were not confounded by response requirements in the MRI scanner, participants performed a no-report version of the task (Fig. 1B). During the initial setup scans, participants were presented with instructions reminding them of the cue-stimulus associations and introducing the catch trial task. To ensure participants continued to pay attention to the stimuli, we introduced a target detection task: participants were asked to detect stimuli in which green pixels were intermixed within the image (face, house or noise). Participants performed 73 practice trials with the first cue-stimulus mapping, lasting ~4 minutes. They then completed 3 × 8-minute blocks of the main task under the first cue-stimulus mapping, each with 145 trials. After each block, the scanner was stopped and participants were asked whether they needed a break. Halfway through the experiment, while their structural scan was obtained, participants were reminded about the second cue-stimulus relationship via another set of instructions and 73 practice

trials. They then completed a further 3 × 8-minute blocks of the main task under the second cue-stimulus mappings. In total, participants performed 870 trials of the main task.

Scanning took place at the Wellcome Centre for Human Neuroimaging, University College London, using a 3 Tesla Siemens Prisma MRI scanner with a 64-channel head coil. We acquired structural images using an MPRAGE sequence (1x1x1 mm voxels, 176 slices, in plane FoV = 256x256mm2), followed by a double-echo FLASH (gradient echo) sequence with TE1 = 10ms and TE2 = 12.46ms (64 slices, slice thickness = 2mm, gap = 1mm, in plane FoV = 192 x 192mm2, resolution = 3 x 3mm2) that was later used for field inhomogeneity correction. Functional scans were acquired using a 2D EPI sequence, optimized for regions near the orbito-frontal cortex (3x3x3mm voxels, TR = 3.36s, TE = 30ms, 48 slices tilted by 30 degrees with respect to the T > C axis, matrix size = 64x72, Z-shim = 1.4).

Model simulations. We used core functions of the Higher-Order State Space (HOSS) model (https://github.com/smfleming/HOSS) to simulate the expected pattern of prediction errors in our experiment (Fleming, 2020). The model is instantiated as a probabilistic graphical model, where nodes correspond to unknown variables and the graph structure indicates dependencies between variables (Fig. 2A). The model is generative, such that higher levels of the hierarchy generate expectations over variables in the layers below. The highest level, the detection (A) state, is a simple scalar such that higher probabilities lead to the activation of content (face or house) states in the W layer below. W is a $1 \times N$ vector that encodes the relative probabilities of each of N discrete perceptual states. Here, $N = 3$, reflecting the 3 possible stimulus categories of face, house or noise. To simulate multivariate sensory data X, we drew samples from one of three multivariate normal distributions conditioned on W: 'noise' with $\mu = [0.5\ 0.5]$, 'face' with $\mu = [1.5\ 0.5]$ or 'house' with $\mu = [0.5\ 1.5]$. The covariance matrix was specified as $\Sigma = [0.1\ 0; 0\ 0.1]$. The locations of samples in evidence space are arbitrary; what is important is the mapping between stimulus categories and the detection state. The likelihood of X given W is then:

$$P(X = x|W) \sim N(\mu_W, \Sigma)$$

Upon receipt of a sample of X, the model can be inverted to compute the posteriors over A and W by marginalising:

$$P(A|X = x) \propto \sum_W P(A)P(W|A)P(X = x|W)$$

$$P(W|X = x) \propto \sum_{A} P(A)P(W|A)P(X = x|W)$$

We simulated four possible prior states, reflecting the four cues in the experiment: $p(A) = 0.8 \,\&\, p(W_{face} = 0.8)$, $p(A) = 0.8 \,\&\, p(W_{face} = 0.2)$, $p(A) = 0.2 \,\&\, p(W_{face} = 0.8)$, $p(A) = 0.2 \,\&\, p(W_{face} = 0.2)$. For each cue-target combination, prediction error was computed at both the detection (A) and content (W) layers as the Kullback-Leibler (KL) divergence between the prior and posterior distributions. The simulated categorisation response was determined by the W-state with the highest poster probability. Only correct trials were used to calculate the prediction errors. We simulated 300 trials per cue-target combination and 30 participants in total.

Behavioural analysis. We first tested whether there was a congruency effect of the content and detection cues by comparing valid versus invalid trials for both cue types using simple t-tests on both accuracy and reaction time (RT). Trials with RTs faster than 200ms or slower than 2s were removed prior to analysis. For the content congruency effects, noise trials were ignored. To investigate the effects of W-level (content) and A-level (detection) prediction errors in more detail, we ran a linear mixed-effects analysis using MATLAB's (R2021b) 'lme.m' function, with the following model:

$$\log(RT) \sim KL_W + KL_A + KL_W \times KL_A + (1 \mid participant)$$

where KL_W and KL_A are proxies for the qualitative patterns expected for W and A-level prediction errors per cue-target combination respectively, calculated using the HOSS model.

fMRI pre-processing. Data pre-processing followed the procedure described in (Mazor et al., 2020; Morales et al., 2018): Imaging analysis was performed using SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm). The first five volumes of each run were discarded to allow for T1 stabilization. Functional images were realigned and unwarped using local field maps (Andersson et al., 2001) and then slice-time corrected (Sladky et al., 2011). Each participant's structural image was segmented into gray matter, white matter, CSF, bone, soft tissue, and air/background images using a nonlinear deformation field to map it onto template tissue probability maps (Ashburner & Friston, 2005). This mapping was applied to both structural and functional images to create normalized images in Montreal Neurological Institute (MNI) space. Normalized images were spatially smoothed using a Gaussian kernel (6 mm FWHM). We set a within-run 4 mm affine motion cut-off criterion. Pre-processing and construction of first- and second-level models used standardized pipelines and scripts available at https://github.com/metacoglab/MetaLabCore/.

Univariate analysis. To test where in the brain activation correlated with W and A-level prediction errors, we performed univariate analyses within SPM12 in MATLAB R2021b. Main effects of A (detection) and W (content) level prediction errors were characterised using the model-predicted KL divergence per trial type (Fig. 2). The general linear model (GLM) contained one regressor aligned to the onset of the stimulus with two parametric modulators, one for each type of prediction error. The onset of the cues and responses were included as nuisance regressors, as were movement parameters, their first derivates and the mean amplitudes of voxels containing white matter and cerebral spinal fluid (CSF). Regressors were specified per run. Significance testing was implemented at the group-level with a t-test of each KL regressor against 0. Correction for multiple comparisons was applied at the cluster-level ($P < 0.05$, family-wise error corrected), using a cluster-forming threshold of $P < 0.001$, uncorrected. Effects were small-volume corrected using either (a) a posterior mask that was generated by combining the following regions from the AAL atlas (Destrieux et al., 2010): all occipital regions, inferior temporal gyrus, calcarine, cuneus, and lingual gyrus or (b) a frontal mask that included: all frontal regions, rectus, insula and anterior cingulate. For the representational similarity analysis (see below), beta weights per cue and target combination were estimated by running a separate GLM with the same nuisance regressors as before but now instead of the KL divergence, including a condition regressor per cue-target combination (12 in total), centred on the target onset.

Representational similarity analysis. To investigate the representational structure of the different types of prediction errors, we performed a searchlight representational similarity analysis (RSA). RSA was performed using MATLAB R2021b and Timo Flesch's RSA toolbox (https://github.com/TimoFlesch/fmri_utils/tree/master/RSA) in combination with custom MATLAB code. We defined 5 different model RDMs (Representational Dissimilarity Matrices) encoding dissimilarity between (1) the presented stimuli, (2) content priors, (3) detection priors, (4) content prediction errors and (5) detection prediction errors (Fig. 5A). Neural RDMs were generated per participant by calculating the Euclidean distance between the activation patterns of different conditions. All within-run comparisons were set to NaN, distances along the diagonals of all model and neural RDMs were also set to NaN, and the lower-triangles were transformed into distance vectors. Per searchlight, a GLM was run to predict the neural RDM from the model RDMs. Group-level inference was performed by testing the inferred beta weights per RDM regressor over participants against 0 using a one-sample t-test. Correction for multiple-comparisons was performed using family-wise error correction ($p < 0.05$) at the whole-brain level, using a cluster-forming threshold of $p < 0.001$, uncorrected (as before).

Results

To independently manipulate predictions about perceptual content and detection of content, we developed a novel perceptual discrimination task with compound cues (Fig. 1). In the behavioural version of the experiment, the task was to infer whether a briefly shown stimulus was a face, a house or noise (Fig. 1A). Preceding the stimulus was a compound cue in which the shape indicated the probability of seeing a stimulus (face or house) versus noise, regardless of its identity – an expectation about the detection of content, rather than content itself. In contrast, the colour of the cue indicated the probability of a stimulus being a face or a house, regardless of whether it was likely to be present – an expectation about content, rather than detection of content. For example, a blue circle indicated that there was a high likelihood no stimulus would be shown (detection expectation) but that if a stimulus was shown, it would likely be a face (content expectation; Fig. 1C).

Simulations predict diverging patterns of prediction error on content and detection.

We used the higher-order state space (HOSS) model (Fleming, 2020) to simulate expected patterns of prediction errors (PEs) for content and detection expectations in our experiment. HOSS specifies a Bayesian network in which higher-order "detection states" (*A*) furnish expectations about the presence of content, with content-specific "content states" (*W*) nested under the detection state layer (Fig. 2A). In the current experiment, content states *W* denote [face, house, noise] and detection state *A* encodes [present, absent] irrespective of content (face and house are mapped to "present", and noise to "absent"). We simulated the model with priors for the *W* and *A* layers set to the empirical prior probabilities used to construct the compound cues. Upon receipt of a multivariate sensory input *X*, the model is inverted and posterior probabilities over both content and detection of content can be derived.

HOSS naturally nests high-dimensional perceptual content (the *W* layer) within a more abstract state that tracks the magnitude or reliability of higher-dimensional perceptual signals. However, HOSS is only one of several possible architectures that could support dissociable inferences on content and detection of content – for instance, a "flat" architecture with no explicit representation of global presence vs. absence may suffice (Whyte & Smith, 2021). The more general point is that inferences on detection of content are proposed to be distinct from inferences on specific contents, with the former being factorised with respect to the latter (Fleming, 2020).

This can be appreciated in the pattern of prediction errors simulated from the model (Figure 2). Simulated prediction errors (Kullback-Leibler divergences) within the detection and content layers for each cue-target combination are shown in Figures 2B and 2C, respectively. In Figure 2B, inferences

about detection of content are predominantly sensitive to expectations about presence vs. absence, and not about specific stimulus contents. For example, a large detection PE is generated when the model expects to see something, but only noise is presented, irrespective of whether the content layer is expecting a face (prFace) or house (prHouse, Fig. 2B; two green bars on the left). In contrast, within the content layer, the largest PEs are observed when content expectations are violated, irrespective of detection expectations. For example, a large PE is generated when a face is expected but a house is presented, regardless of whether the detection layer expected to see a stimulus (prFace) or not (abFace, Fig. 2C; dark blue bars). Note that in the content layer, prediction errors for noise are low because a noise patch contains an absence of information about either feature (face or house). Strictly, in the model, noise is an absence of any input supporting either of the two features – whereas in our experimental paradigm, we use a noise patch to indicate the 'absence' of content. As noise does not contain meaningful content at the level of object categories such as faces and houses, this is a reasonable approximation to the simulations. Crucially, the shared variance between the A and W PEs is low (correlation of -0.14 between the two simulated condition vectors), showing that our experimental design allows us to independently investigate neural and behavioural correlates of content and detection PEs.

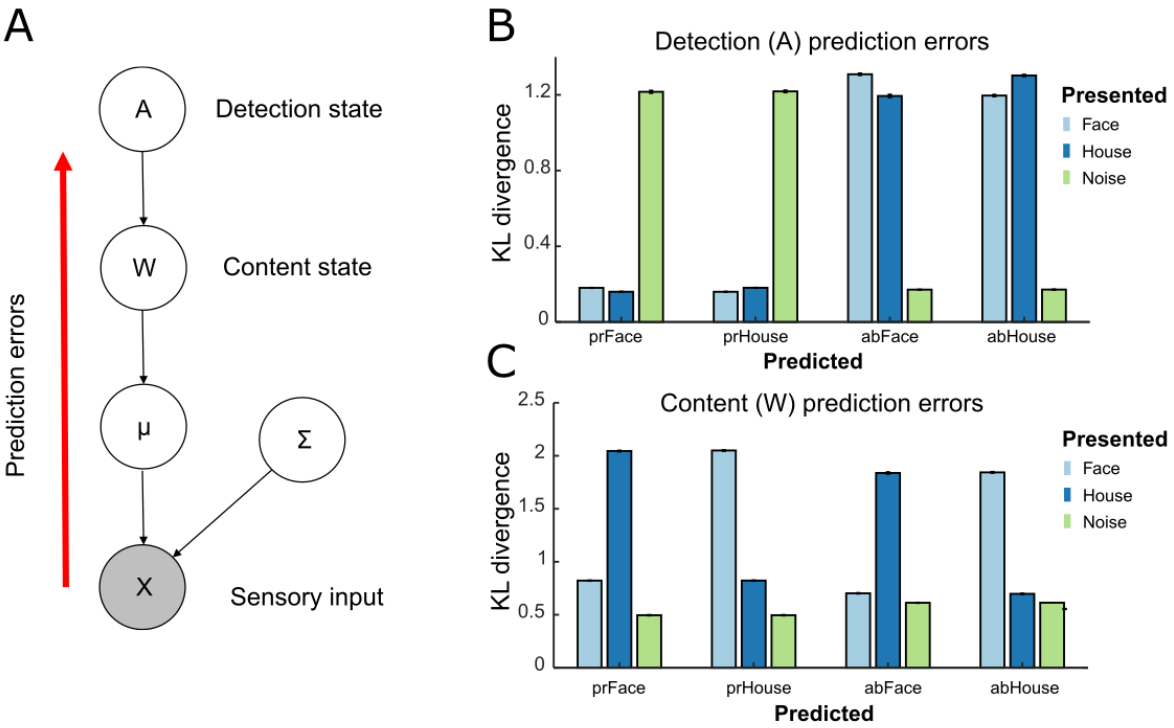

Figure 2. Simulated prediction errors within content and detection layers of the higher-order state space (HOSS) model. (**A**) Graphical representation of the HOSS model. Perceptual states W and detection state A are inferred based on sensory input X. Simulated prediction errors per cue-target combination within the A-detection layer (**B**) and W-content layer (**C**) are plotted. The x-axis reflects the cues, with the first two letters indicating whether the cue indicated high probability of the presence (pr) or absence (ab) of content (the

presence expectation), followed by whether it indicated a high probability of being face or house (the content expectation). The y-axis indicates the simulated KL-divergence between the prior and the posterior at the different levels of the model (a proxy for prediction error, or how much belief change is induced by a sensory sample). Error bars indicates standard errors of the mean (SEM) over simulation samples.

Content and detection expectations both influence behaviour.

We next investigated whether detection and content expectations separately modulated discrimination reaction times in a behavioural experiment. The logic of our approach is grounded in previous findings that expected stimuli lead to faster responses, whereas violations of expectations (prediction errors) lead to behavioural slowing (Brodersen et al., 2008; Carpenter & Williams, 1995; Mars et al., 2008). Thirty-six participants performed the behavioural version of the task (Fig. 1A). We first ran a model-free analysis to investigate whether there were congruency effects of content (only face and house trials) and detection cues on reaction time (RT). RTs were indeed faster for trials with a valid content cue (M = 631.38ms, SD = 101.63ms) compared to trials with an invalid content cue (M = 644ms, SD = 109ms; $t(35) = 2.12$, $p = 0.041$, $d = 0.12$) and also faster for trials with a valid detection cue (M = 629ms, SD = 109ms) compared to trials with an invalid presence cue (M = 638ms, SD = 104ms, $t(35) = 2.04$, $p = 0.049$, $d = 0.08$). The interaction between the two cue dimensions (content and detection) was not significant ($t(35) = -1.92$, $p = 0.063$).

We next ran a model-based linear mixed-effects regression analysis predicting reaction times from the simulated content and detection PEs obtained from the HOSS model (entered as random effects) with random intercepts for each participant (Fig. 3B). Both content PEs ($beta$ = 0.028, $t(16508.07) = 3.01$, $p = 0.0026$; Fig. 3B, blue bar) as well as detection PEs ($beta$ = 0.022, $t(16508.04) = 2.21$, $p = 0.027$; Fig. 3B, red bar) led to significant increases in reaction time. The interaction between content and detection PEs was again not significant ($t(16508.04) = -0.63$, $p = 0.53$; Fig. 3B, grey bar).

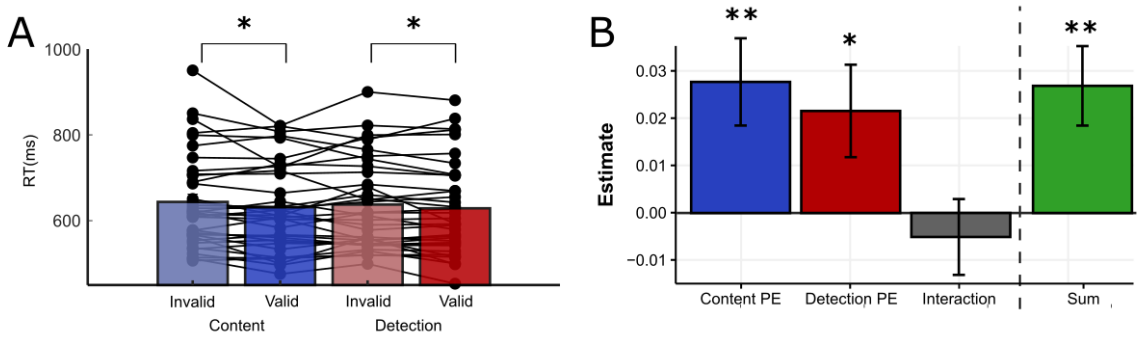

Figure 3. Behavioural results. (**A**) Reaction times (RTs) separated by valid and invalid expectations on content (blue) and detection (red). Points represent individual participants (**B**) Results of a linear mixed-effects analysis using simulated prediction errors (Fig. 2) as predictors of response times. The green bar reflects the beta estimate for a model including a predictor consisting of the sum of the simulated content and detection

prediction errors on each trial. * $p < 0.05$, ** $p < 0.005$, *** $p < 0.0005$. Error bars reflect standard errors of the mean (SEM).

To explore whether the effects of content and detection PEs on reaction times were additive, we also tested another model in which the PEs of the two components were summed (i.e. the sum of the K-L divergences displayed in Fig. 2A and B) to create one combined PE regressor per trial. This summed predictor was associated with a significant increase in reaction time, as expected (*beta* = 0.027, $t(16541) = 3.19$, $p = 0.0014$; Fig. 3B, green bar). Model comparison indicated that a model containing the summed PE (BIC = 417.81) was a more parsimonious explanation of the reaction time data than a model with independent PE terms (BIC = 426.23).

Taken together, our analysis of reaction time effects reveals that, at the level of behaviour, detection and content PEs both lead to significant slowing, and that their influence is best modelled as a linear sum of the two PE terms. On the basis of behavioural data alone, we are unable to conclude whether inferences on content and detection of content are supported by distinct (neural) computations. Therefore, in order to investigate this, we next turned to neuroimaging to ask how these two types of PEs are encoded in the brain.

Neural correlates of content and detection prediction errors.

Twenty-seven of the participants who had completed the behavioural experiment went on to perform a no-report version of the same task while undergoing whole-brain functional neuroimaging (Fig. 1B). Participants passively viewed the compound cues and stimuli without being required to explicitly categorise each stimulus. Instead, to ensure attention, a button press was required on catch trials (indicated by green pixels within the stimulus, 20% trials). Given that the catch manipulation was independent of our main effects of interest, these trials were included in subsequent analyses.

We pursued two complementary analysis approaches that aimed to identify a) univariate signals and b) multivariate patterns covarying with either content or detection PEs. First, to identify univariate brain activity modulated by detection and content PEs, we ran a whole-brain GLM entering the simulated K-L divergences from the detection and content layers of the HOSS model as regressors. Content and detection PEs correlated with activation in different brain areas (Fig. 4). To test our a priori hypotheses that content PEs would be observed in visual sensory areas and detection PEs in prefrontal areas, we applied small-volume corrections based on posterior and frontal masks.

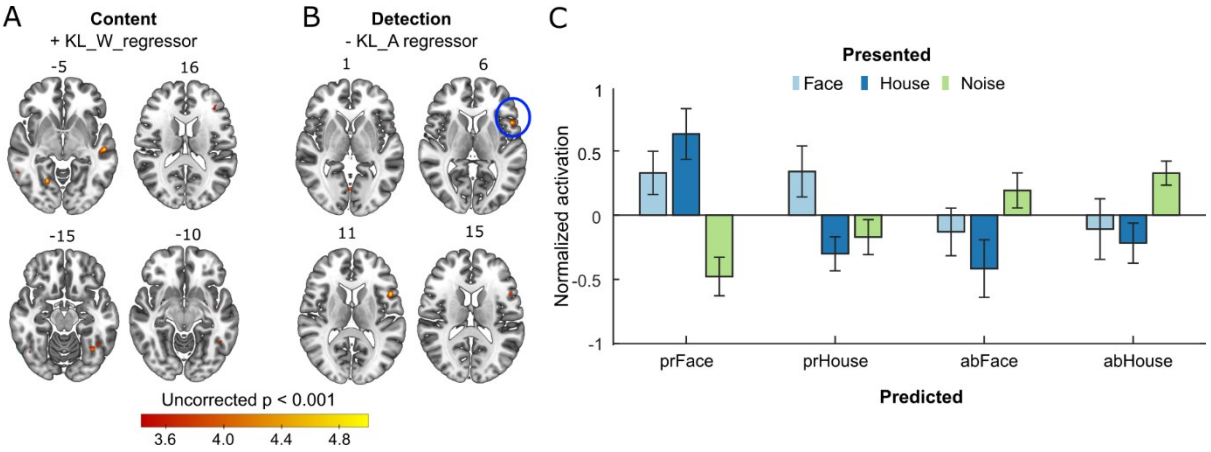

Figure 4. Univariate correlates of prediction errors on detection and content. (**A,B**) Brain areas that significantly correlated positively with the KL W regressor (**A**) and negatively with the KL A regressor (**B**) thresholded at p < 0.001 uncorrected. (**C**) Activation profile of the region that showed a significant (p<0.05, FWE-small-volume corrected) effect of the KL A regressors, circled in blue in **(B)**. Activation is z-scored per participant to account for large variations in mean amplitude between participants. Error bars reflect SEM.

A positive effect of the content PE regressor in the fusiform gyrus (Fig. 4A) did not survive correction for multiple comparisons and we therefore refrain from interpreting it further (t(26) = 4.77, *cluster-level* $p_{\text{FWE-corrected}}$ = 0.052). The detection PE regressor showed a *negative* correlation with activation in the left inferior frontal cortex when applying small-volume correction within a frontal mask (IFC; *t*(26) = 5.55, *cluster-level* $p_{\text{FWE-corrected}}$ = 0.026; Fig. 4C), close to voxels showing (uncorrected) effects of the content PE regressor. There was no significant detection PE effect when applying small-volume correction within a posterior mask (all p-values > 0.338). Within this region, when presence was expected (prFace and prHouse) activation tended to be higher when this expectation was confirmed and content was presented (irrespective of whether this stimulus was a face or a house) compared to when noise was presented (Fig. 4D). Conversely, when no content was expected (abFace or abHouse), activation was higher when this expectation was confirmed, and noise was presented, compared to when a face or house was presented (Fig. 4D). These results show that activity in left IFC decreases when detection predictions are violated.

Rich representations of content and detection prediction errors in sensory and prefrontal brain networks.

While univariate analyses can reveal low-dimensional neural signatures of content and detection prediction errors, they are blind to changes in distributed neural codes that might support prior and posterior beliefs in richer representational spaces. Therefore, we next investigated whether prediction errors were also encoded in multivariate patterns rather than univariate amplitude differences, using representational similarity analysis (RSA). We formulated representational dissimilarity matrices

(RDMs) that identified dissimilarities between conditions in either the prior (cue), stimulus or prediction error, with prior and prediction error RDMs being specified separately for the content and detection layers of the model (Fig. 5A). These RDMs were designed to test whether the activity patterns in one set of conditions sharing a specific feature (e.g. all conditions in which a face was presented) were more similar to each other (lower distance between them) than to conditions with a different feature.

For the prediction error RDMs, we hypothesized that if an area coded for PEs at a given level, the presence of a PE would lead to convergence towards a specific activity pattern. In contrast, in the absence of a PE at that level, activity patterns would reflect noise, and be random (uncorrelated). Therefore, for both content and detection PEs, we assumed that conditions with a PE would be similar to other conditions with a PE, whereas conditions without a PE would be dissimilar both to each other and to conditions with a PE (Fig. 5A). Note that for all RSA analyses, the diagonal elements were removed before computing similarity with neural data.

We found strong positive correlations between the stimulus RDMs and posterior brain regions (Fig. 5B, Appendix A - Table 1) indicating that within these regions, stimuli belonging to the same category were encoded in similar activation patterns, irrespective of prior expectations. Furthermore, we found that the detection prior RDM showed significant positive correlations with activity patterns in the ventromedial prefrontal cortex (vmPFC; Fig. 5C, Appendix A - Table 2), indicating that in this region, the activity patterns of conditions in which presence was expected were more similar to each other than to conditions in which stimulus absence was expected, and vice-versa, irrespective of expectations about content (face or house). We did not find any significant correlation with the content-level prior RDM.

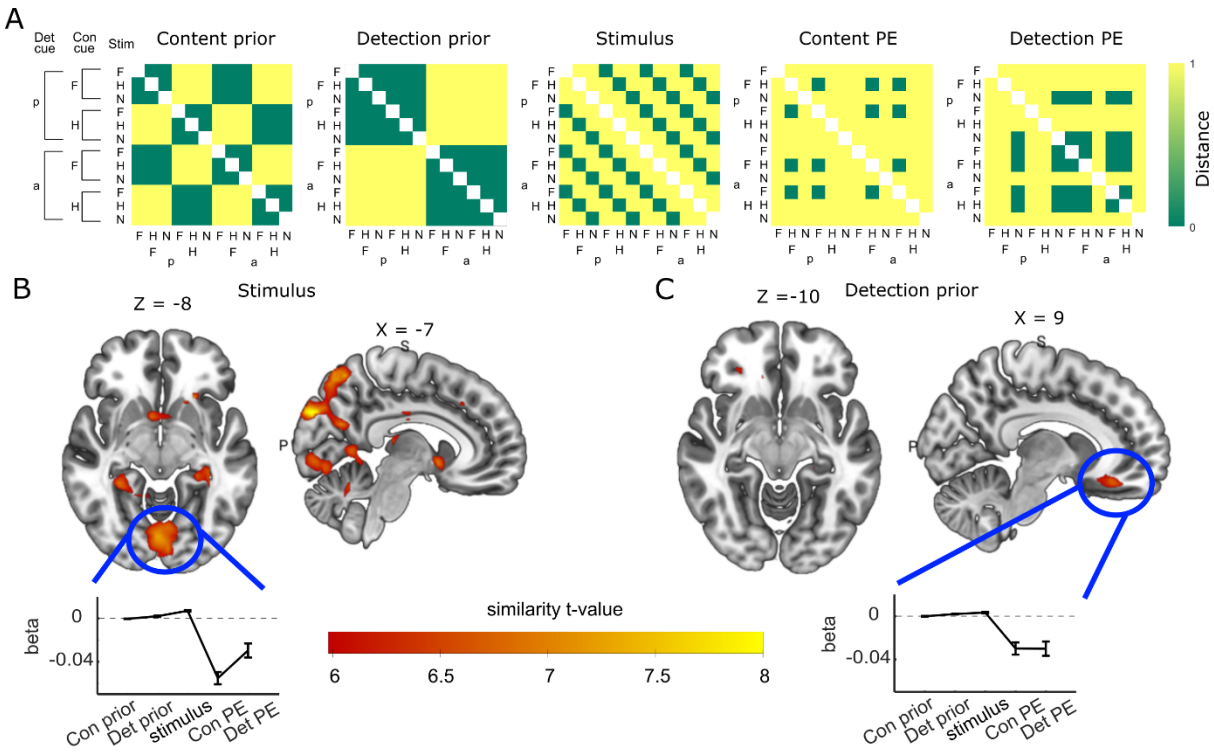

Figure 5. RDM hypotheses and similarity coding of stimulus- and prior-related information. (**A**) Representational Dissimilarity Matrices (RDMs) reflecting different hypotheses about the similarity of neural patterns of different conditions based on encoding of content prior, detection prior, stimulus, content prediction error and detection prediction error. Darker colours indicate higher similarity, i.e. lower distance, between conditions. Only the Stimulus and Detection Prior RDMs showed significant positive correlations with brain activity. (**B**) Similarity in stimulus category encoding (FWE-corrected at $p < 0.05$): conditions in which stimuli of the same category were presented were more similar than those in which different stimuli were presented. Beta values for region of interest (ROI) based on this contrast encircled in blue are shown below. All RDMs significantly predicted activation patterns in this ROI, except the content prior, with all p-values < 0.006 (uncorrected). (**C**) Similarity in detection prior encoding (FWE-corrected at $p < 0.05$): conditions in which presence was expected were more similar to each other than to conditions in which absence was expected and vice-versa. Beta values for the blue encircled ROI are shown below. All RDMs significantly predicted activation patterns in this ROI, expect the content prior, with all p-values < 0.0006 (uncorrected). Statistical maps in (B) and (C) are thresholded at $p < 0.001$ uncorrected; see Tables 2 and 3 for details of clusters surviving whole-brain correction including the EVC and vmPFC ROIs.

When inspecting the beta weights for each of the five RDMs in selected brain regions, we observed that activity patterns in both early visual cortex (EVC) as well as vmPFC showed strong *negative* relationships with both our hypothesized content and detection PE RDMs (see Fig. 5 beta plots below panels B and C). Negative relationships indicate that the conditions that are hypothesized to show similar patterns of activity in fact show dissimilar patterns of activity in the brain, and vice versa. Within the RSA literature, both similarity- and dissimilarity-based neural coding schemes have been hypothesized (Kriegeskorte et al., 2008). For example, in the fusiform face area, faces are encoded as more similar to other faces than to houses (positive correlation with a category membership RDM) whereas in face-selective parts of the inferotemporal cortex (IT), faces are encoded as more *dissimilar* to other faces, indicating exemplar encoding, revealing that this region is sensitive to the identity of

the face (negative correlation with a category membership RDM (Kriegeskorte et al., 2008). In the current context, negative correlations with the hypothesized PE RDMs would therefore be in line with the corresponding brain region being sensitive to the identity of the PE (e.g. whether it tracks a violation of face or house predictions).

Therefore, to further explore these negative correlations, we expanded our search to examine negative correlations with content and detection PE RDMs in a whole-brain searchlight analysis. This approach indeed revealed significant dissimilarity PE encoding in several brain areas (Fig. 6). Content PE-related patterns were predominant in posterior sensory regions (Fig. 6A, Appendix A – Table 3) and detection PE-related patterns were predominant in prefrontal regions (Fig. 6B, Appendix A -Table 4). However, directly comparing detection PE RDMs against content PE RDMs in a whole-brain contrast revealed no significant differences between the two PE maps. Together these results indicate that while neural patterns associated with content and detection PEs are predominantly expressed in different parts of the brain, this distinction is graded rather than discrete.

To further characterize the relationship between content and detection PEs within different brain areas, we examined activity patterns within functional ROIs selected for their dominance of content- or detection-PE effects: the early visual cortex (EVC) and the Precuneus for content PEs and the ventromedial prefrontal cortex (vmPFC) for detection PEs (encircled in blue in Fig. 6A&B). Note that effects of the other RDMs in these ROIs cannot be explained by collinearity between the RDMs, as the maximum correlation between regressors was low (0.11). Besides the expected content PE effect, both the EVC and the Precuneus also showed a significant negative correlation with the detection PE RDM (EVC: $t(26) = -3.68$, $p = 0.0011$; Precuneus: EVC: $t(26) = -4.93$, $p = 0.0008$; Fig 6C&E). Furthermore, the vmPFC ROI defined based on the detection PE effect also showed a significant negative correlation with the content PE RDM ($t(26) = -3.99$, $p = 0.0005$). This suggests that the brain regions that are modulated by content PEs are also sensitive to detection PEs and vice versa.

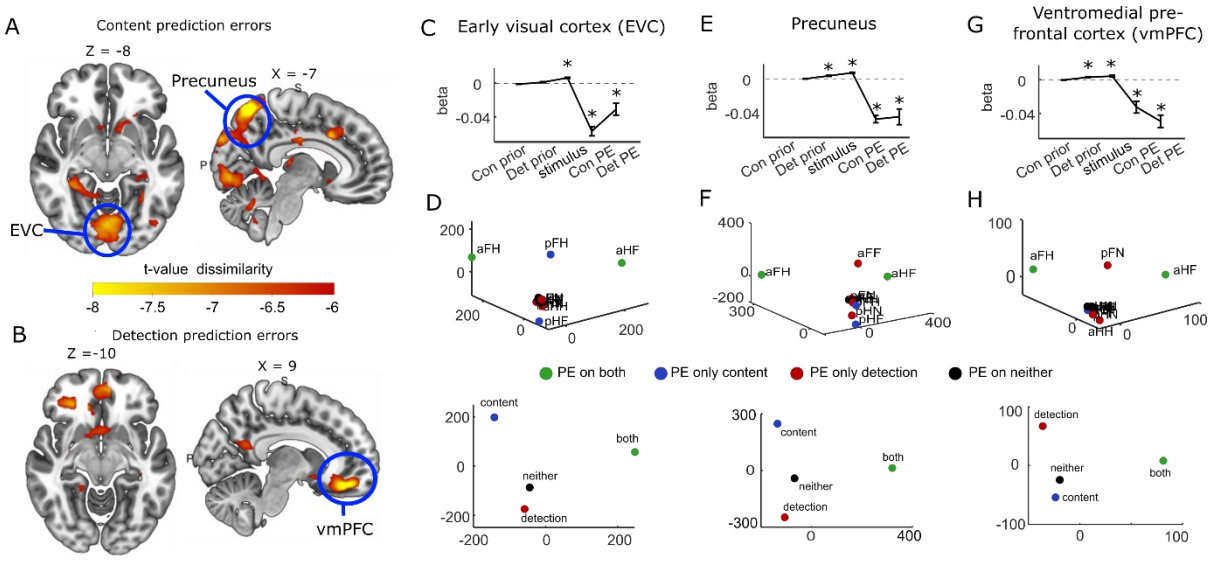

Figure 6. Dissimilarity coding of prediction errors. (**A**) Dissimilarity in the encoding of content prediction errors: conditions in which there was a prediction error on content (e.g. a face was expected and a house was presented) were more dissimilar/less similar to each other than to conditions without a prediction error on content. (**B**) Dissimilarity in the encoding of detection prediction errors: conditions in which there was a prediction error on detection (e.g. noise was expected but a face or house was presented) were more dissimilar/less similar to each other than to conditions without a prediction error on detection. (**C**, **E**, **G**) Beta values for the three blue encircled regions of interest (ROI) for each of the RDMs. * indicates the RDM significantly predicted activation pattern in this ROI (Bonferroni corrected). (**D, F, H**) Multidimensional scaling (MDS) of activation patterns for each cue-target combination in the three ROIs. Top: MDS in 3 dimensions with each cue-target combination plotted separately. Bottom: MDS in 2 dimensions in which conditions are grouped together based on whether they contain prediction errors on both content and detection (green), only on content (blue), only on detection (red) or on neither (black). Statistical maps in (A) and (B) are thresholded at $p < 0.001$ uncorrected; see Tables 4 and 5 for details of clusters surviving whole-brain correction including the EVC, precuneus and vmPFC ROIs.

To further characterize the representational structure of activity profiles in these regions, we performed classical multi-dimensional scaling (MDS) on the similarity of the activation patterns. MDS visualizes the similarity in neural patterns between conditions by projecting the data into a lower dimensional space in which similar conditions are plotted nearby to each other. We note that such visualisations were expected to recapitulate the RDM used to identify the ROI – for instance, we expect that within vmPFC we should see detection PEs as being encoded distinct from other trial types. However, it is possible that the nature of voxel patterns within these candidate ROIs diverge in other interesting ways.

Within the EVC, all conditions that did not contain a content PE were represented as similar to each other (Fig. 6D, red and black dots) whereas conditions containing a content PE were dissimilar to each other as well as to other conditions, as expected from the content PE RDM (Fig. 6C, green and blue dots). Interestingly, however, and in line with the observation that the detection PE RDM also showed a significant effect in this region, content PEs that also violated detection predictions (green dots) were dissimilar both to each other and to prediction errors on content only (blue dots). Together,

this suggests that in the EVC, once a content prediction is violated, activation diverges according to (a) the exact type of content violation (face > house or house > face) and (b) whether the detection prediction is also violated or not. Note that in the EVC, prediction errors on detection only (red dots) were represented as similar to conditions in which no prediction error occurred (black dots), suggesting that the EVC is only modulated by detection PEs when content predictions are violated.

Within the vmPFC, in contrast, all conditions that did not contain a detection PE were represented as similar to each other (Fig. 6F, blue and black dots) whereas conditions containing a detection PE were dissimilar to each other as well as to other conditions, as expected from the detection PE RDM (Fig. 6F, green and red dots). However, and going beyond the hypothesis RDM, MDS showed that trials that additionally violated content predictions (green dots) were also dissimilar to each other and to prediction errors on detection only (red dots), in line with the observation that content PEs also modulate activity patterns in this region. Note that in the vmPFC, in contrast to the EVC, conditions with prediction errors on content only (blue dots) were represented as similar to conditions in which no prediction error occurred (black dots), suggesting the vmPFC is only modulated by content PEs when detection predictions are violated as well. Finally, the (pre-)cuneus showed a significant effect of both the content and detection PE RDMs and closer inspection of the activity patterns in this region (Fig. 6 H) showed that, in contrast to the EVC and vmPFC, in the (pre-)cuneus all conditions with any kind of PE were associated with diverging activity patterns.

Discussion

In this study we set out to test whether inferences about perceptual content (*what* is perceived) and inferences about detection of that content (*whether* something is perceived) are underpinned by distinct neural substrates. To this end, we developed a novel experimental paradigm that used compound cues to separately induce expectations about perceptual content and detection of that content. We found that both content and detection expectations influenced reaction times, with higher probability stimuli being identified more quickly. Using a no-report version of this paradigm in conjunction with fMRI, we found that prediction errors on content correlated most strongly with posterior visual brain areas, whereas prediction errors on detection correlated most strongly with prefrontal brain areas. However, contrary to our hypothesis, these representations were not orthogonal. Instead, prediction errors on one level gated the expression of prediction errors at the other level. Taken together, our results suggest that inferences on content and detection of content rely on distinct but interacting neural computations.

We observed a negative univariate effect of detection prediction errors in IFC, indicating that in this region, activation was higher when a detection prediction was confirmed. One possibility is that

this region encodes a detection prior, with activity being strengthened in conditions in which these priors are reinforced by matching input. However, within most neuronal models of predictive processing, priors and prediction errors are assumed to be encoded within the same brain region (Bastos et al., 2012). An alternative possibility is that a confirmation effect in IFC reflects a signature of perceptual confidence (Cortese et al., 2016; Hilgenstock et al., 2014; Shekhar & Rahnev, 2018), which is likely to be higher when predictions are confirmed compared to when they are violated. Further work is needed to identify neural substrates supporting putative detection-specific confidence signals, and distinguish these from other aspects of metacognition (Mazor et al., 2020, 2022).

Furthermore, our representational similarity analyses (RSA) revealed that prediction errors at both content and detection levels were encoded as being dissimilar to each other. The pattern of dissimilarity that we found indicates a sharp distinction in the initial trigger for PE coding in EVC and prefrontal cortex – in EVC, the trigger for coding PEs is a violation of content expectations, whereas in vmPFC, the trigger for coding PEs is a violation of detection expectations. Specifically, a divergence in activity patterns in EVC is triggered by a content PE – here, a violation in the expectation of face or house. Once a content PE is triggered, then EVC activity patterns go on to represent the type of PE within the full compound cue space, tracking violations of both content and detection expectations. In contrast, a divergence in activity patterns in vmPFC is triggered by a detection PE – whether an absence expectation has been violated by stimulus presence, or whether a presence expectation has been violated by stimulus absence. Once a detection PE is triggered, then vmPFC activity patterns also go on to represent the type of PE within the full compound cue space, inheriting information about content (face vs. house) violations. The Precuneus shows an intermediate effect, representing diverging patterns for every type of PE. This pattern dissimilarity suggests that prediction errors are encoded in an exemplar specific way, similar to individual faces in IT (Kriegeskorte et al., 2008). Further work is needed to fully understand what this dissimilarity-based coding implies for the computational underpinnings of inferences about perceptual content and detection.

As our neuroimaging results were obtained in the absence of reports, they provide evidence in favour of an architecture in which detection prediction errors are automatically elicited even under passive viewing conditions. Our results are consistent with other studies observing prefrontal correlates of subjective detection in the absence of overt report (Hatamimajoumerd et al., 2022). We note that the focus of the current study is on distinguishing between neural signatures of inference on content, and detection of content, and did not set out to measure variation in subjective perception or awareness. However, our results bear on the possible neural architectures supporting predictive processing accounts consciousness, as inferences on detection – i.e. whether subjects are “aware” or “unaware” of particular stimulus features – are the cornerstone of conscious reportability. We also

note that our findings cannot be explained by mere stimulus effects, as our focus here is on how the stimulus interacts with an experimentally-manipulated expectation, thereby generating a prediction error signal. Future studies are necessary to investigate whether the inferences on content and detection we identify here relate to changes in conscious experience while keeping stimulus input near threshold (Frith et al., 1999; Leopold & Logothetis, 1996).

One example of such an approach is a recent MEG study which revealed a neural signature of the content of false percepts in the occipital lobe, whereas confidence in stimulus detection was reflected in a parieto-frontal network (Haarsma, Hetenyi, et al., 2024). Combining such a false percept paradigm with the type of compound content-detection cue employed here (Haarsma, Kaltenmaier, et al., 2024) is a promising avenue to investigate how the neural correlates identified here relate to fluctuations in conscious experience. To ensure that effects are not due to reports in that case, future research could develop no-report read-outs of perceptual content such as eye movements (Frassle et al., 2014; Frässle et al., 2013) to capture aspects of both discrimination and detection. Finally, It would also be interesting to seek to causally intervene on regions (such as vmPFC) exhibiting signatures of detection prediction error (for instance, using multivariate neurofeedback (Taschereau-Dumouchel et al., 2021)), and ask whether and how such interventions alter conscious experience.

Several ideas have been advanced to accommodate conscious awareness (in the form of subjective detection) within a predictive processing framework (Clark et al., 2019; Doerig et al., 2020; Fleming, 2020; Hobson & Friston, 2014, 2012; Hohwy et al., 2008; Hohwy & Seth, 2020). One instantiation of such an architecture proposes that conscious detection arises from inferences deep within a perceptual hierarchy. Interestingly, in line with this hierarchical view, detection PEs were preferentially localised to a vmPFC region overlapping with the default model network (DMN) (Raichle, 2015) which is proposed to occupy a deep position within a cognitive hierarchy (Margulies et al., 2016). Other work has linked the vmPFC to carrying information about latent (unobservable) perceptual spaces, such as hidden states governing task structure, or links between arbitrary stimuli on a graph (Park et al., 2020; Schuck et al., 2016). Another key node of the DMN, the Precuneus, was also evident in the PE RSA analysis, and showed clear detection- as well as content-related prediction error effects (Raichle, 2015). Other research has shown that the Precuneus is modulated by both the level of awareness (Bisenius et al., 2015; Cavanna, 2007; Kjaer et al., 2001) and represents stimulus content (Doesburg et al., 2009) – consistent with it inhabiting an intermediate position in a perceptual hierarchy.

In conclusion, using a novel experimental paradigm we show that prediction errors on perceptual content and detection of content are encoded in distinct but interacting activity patterns in the human brain. These results are consistent with a proposal that detection may require distinct

neural computations that go beyond those required for inferences on content itself. More generally, our findings provide a framework for future empirical and theoretical studies that incorporate and model detection and discrimination as distinct dimensions within powerful predictive processing accounts of perception and cognition.

Data Availability

Second-level maps are uploaded on NeuroVault at: https://neurovault.org/collections/14778/.

Code Availability

All analysis code is available on: https://github.com/NadineDijkstra/HSFPA

Author contribution statement

Conceptualization: N.D., P.K. & S.M.F.; Data Curation: N.D. & O.W.; Formal Analysis: N.D. & O.W.; Funding Acquisition: S.M.F.; Investigation: N.D. & O.W.; Methodology: N.D., O.W. & S.M.F.; Project Administration: N.D. & O.W.; Resources: S.M.F.; Software: N.D. & O.W.; Supervision: N.D. & S.M.F.; Validation: N.D.; Visualization: N.D. & O.W.; Writing – original draft: N.D.; Writing – review & editing: N.D., O.W., P.K. & S.M.F.

Gender citation bias

Number of DOIs categorized: 54, Number of DOIs due to missing author data: 7. Proportion per category: MM 0.87; WM: 0.074; MW: 0.037; WW: 0.019. GCBI per category: MM 1.139; WM -0.769; MW -0.678; WW: -0.884.

References

Andersson, J. L. R., Hutton, C., Ashburner, J., Turner, R., & Friston, K. (2001). Modeling Geometric Deformations in EPI Time Series. *NeuroImage*, *13*(5), 903–919. https://doi.org/10.1006/NIMG.2001.0746

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *NeuroImage*, *26*(3), 839–851. https://doi.org/10.1016/J.NEUROIMAGE.2005.02.018

Azzopardi, P., & Cowey, A. (1997). Is blindsight like normal, near-threshold vision? *Proceedings of the National Academy of Sciences of the United States of America*, *94*(25). https://doi.org/10.1073/pnas.94.25.14190

690 Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical
691 Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711.
692 https://doi.org/10.1016/j.neuron.2012.10.038
693 Bisenius, S., Trapp, S., Neumann, J., & Schroeter, M. L. (2015). Identifying neural correlates of visual
694 consciousness with ALE meta-analyses. *NeuroImage*, *122*, 177–187.
695 https://doi.org/10.1016/j.neuroimage.2015.07.070
696 Brodersen, K. H., Penny, W. D., Harrison, L. M., Daunizeau, J., Ruff, C. C., Duzel, E., Friston, K. J., &
697 Stephan, K. E. (2008). Integrated Bayesian models of learning and decision making for
698 saccadic eye movements. *Neural Networks*, *21*(9).
699 https://doi.org/10.1016/j.neunet.2008.08.007
700 Brown, R. (2015). The HOROR theory of phenomenal consciousness. *Philosophical Studies*, *172*(7),
701 1783–1794. https://doi.org/10.1007/s11098-014-0388-7
702 Carpenter, R. H. S., & Williams, M. L. L. (1995). Neural computation of log likelihood in control of
703 saccadic eye movements. *Nature*, *377*(6544). https://doi.org/10.1038/377059a0
704 Cavanna, A. E. (2007). The Precuneus and Consciousness. *CNS Spectrums*, *12*(7), 545–552.
705 https://doi.org/10.1017/S1092852900021295
706 Clark, A., Friston, K., & Wilkinson, S. (2019). Bayesing qualia: Consciousness as inference, not raw
707 datum. *Journal of Consciousness Studies*, *26*(9–10), 19–33.
708 Cortese, A., Amano, K., Koizumi, A., Kawato, M., & Lau, H. (2016). Multivoxel neurofeedback
709 selectively modulates confidence without changing perceptual performance. *Nature*
710 *Communications*, *7*(1), 1–18. https://doi.org/10.1038/ncomms13669
711 Dehaene, S., & Changeux, J. P. (2011). Experimental and Theoretical Approaches to Conscious
712 Processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018
713 Dehaene, S., Naccache, L., Cohen, L., Bihan, D. L., Mangin, J. F., Poline, J. B., & Rivière, D. (2001).
714 Cerebral mechanisms of word masking and unconscious repetition priming. *Nature*
715 *Neuroscience*, *4*(7), 752–758. https://doi.org/10.1038/89551

716 Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri
717 and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15.
718 https://doi.org/10.1016/j.neuroimage.2010.06.010
719 Doerig, A., Schurger, A., & Herzog, M. H. (2020). Hard criteria for empirical theories of consciousness.
720 *Cognitive Neuroscience*, 1–22. https://doi.org/10.1080/17588928.2020.1772214
721 Doesburg, S. M., Green, J. J., McDonald, J. J., & Ward, L. M. (2009). Rhythms of consciousness:
722 Binocular rivalry reveals large-scale oscillatory network dynamics mediating visual
723 perception. *PloS One*, *4*(7), e6142. https://doi.org/10.1371/journal.pone.0006142
724 Fleming, S. M. (2020). Awareness as inference in a higher-order state space. *Neuroscience of*
725 *Consciousness*, *2020*(1). https://doi.org/10.1093/nc/niz020
726 Frassle, S., Sommer, J., Jansen, A., Naber, M., & Einhauser, W. (2014). Binocular Rivalry: Frontal
727 Activity Relates to Introspection and Action But Not to Perception. *Journal of Neuroscience*,
728 *34*(5), 1738–1747. https://doi.org/10.1523/JNEUROSCI.4403-13.2014
729 Frässle, S., Sommer, J., Naber, M., Jansen, A., & Einhäuser, W. (2013). Neural Correlates of Binocular
730 Rivalry as measured in fMRI are partially confounded by observers' active report. *Journal of*
731 *Vision*, *13*(9), 937. https://doi.org/10.1167/13.9.937
732 Friston, K., Kilner, J., & Harrison, L. (2006). A free energy principle for the brain. *Journal of Physiology*
733 *Paris*, *100*(1–3), 70–87. https://doi.org/10.1016/j.jphysparis.2006.10.001
734 Frith, C., Perry, R., & Lumer, E. (1999). The neural correlates of conscious experience: An
735 experimental framework. *Trends in Cognitive Sciences*, *3*(3), 105–114.
736 https://doi.org/10.1016/s1364-6613(99)01281-4
737 Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. In *John Wiley* (Vol. 5).
738 https://doi.org/10.1016/0022-460x(67)90197-6
739 Haarsma, J., Hetenyi, D., & Kok, P. (2024). *Shared and diverging neural dynamics underlying false and*
740 *veridical perception* (p. 2023.11.16.567367). bioRxiv.
741 https://doi.org/10.1101/2023.11.16.567367

Haarsma, J., Kaltenmaier, A., Fleming, S. M., & Kok, P. (2024). *Expectations about presence enhance the influence of content-specific expectations on low-level orientation judgements* (p. 2024.02.22.581334). bioRxiv. https://doi.org/10.1101/2024.02.22.581334

Hatamimajoumerd, E., Ratan Murty, N. A., Pitts, M., & Cohen, M. A. (2022). Decoding perceptual awareness across the brain with a no-report fMRI masking paradigm. *Current Biology*, *32*(19), 4139-4149.e4. https://doi.org/10.1016/J.CUB.2022.07.068

Hilgenstock, R., Weiss, T., & Witte, O. W. (2014). You'd Better Think Twice: Post-Decision Perceptual Confidence. *NeuroImage*, *99*. https://doi.org/10.1016/j.neuroimage.2014.05.049

Hobson, J. A., & Friston, K. (2014). Consciousness, dreams, and inference: The cartesian theatre revisited. *Journal of Consciousness Studies*, *21*(1–2), 6–32.

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology*, *98*(1), 82–98. https://doi.org/10.1016/J.PNEUROBIO.2012.05.003

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology*, *3*(April), 96. https://doi.org/10.3389/fpsyg.2012.00096

Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. https://doi.org/10.1016/j.cognition.2008.05.010

Hohwy, J., & Seth, A. (2020). Predictive processing as a systematic basis for identifying the neural correlates of consciousness. *Philosophy and the Mind Sciences*, *1*(II). https://doi.org/10.33735/phimisci.2020.ii.64

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

766 Kjaer, T. W., Nowak, M., Kjaer, K. W., Lou, A. R., & Lou, H. C. (2001). Precuneus-prefrontal activity
767 during awareness of visual verbal stimuli. *Consciousness and Cognition*, *10*(3), 356–365.
768 https://doi.org/10.1006/ccog.2001.0509
769 Kok, P., Brouwer, G. J., van Gerven, M. A. J., & de Lange, F. P. (2013). Prior expectations bias sensory
770 representations in visual cortex. *Journal of Neuroscience*, *33*(41).
771 https://doi.org/10.1523/JNEUROSCI.0742-13.2013
772 Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting
773 the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4.
774 https://doi.org/10.3389/neuro.06.004.2008
775 Lau, H. (2019). Consciousness, Metacognition, & Perceptual Reality Monitoring. *PsychArxiv*, 1–17.
776 Lau, H. (2022). In Consciousness we Trust. In *In Consciousness we Trust*.
777 https://doi.org/10.1093/oso/9780198856771.001.0001
778 Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. *Progress in Brain*
779 *Research*, *168*, 35–48. https://doi.org/10.1016/S0079-6123(07)68004-2
780 Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness.
781 *Trends in Cognitive Sciences*, *15*(8), 365–373. https://doi.org/10.1016/j.tics.2011.05.009
782 Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys'
783 percepts during binocular rivalry. *Nature*, *379*, 549–553. https://doi.org/10.1038/379549a0
784 Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces—KDEF, CD*
785 *ROM from Department of Clinical Neuroscienc*.
786 Marcel, A. J. (1983). Conscious and unconscious perception: Experiments on visual masking and
787 word recognition. *Cognitive Psychology*, *15*(2). https://doi.org/10.1016/0010-
788 0285(83)90009-9
789 Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G.,
790 Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating
791 the default-mode network along a principal gradient of macroscale cortical organization.

Proceedings of the National Academy of Sciences of the United States of America, *113*(44). https://doi.org/10.1073/pnas.1608282113

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., & Bestmann, S. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, *28*(47). https://doi.org/10.1523/JNEUROSCI.2925-08.2008

Marvan, T., & Havlík, M. (2021). Is predictive processing a theory of perceptual consciousness? *New Ideas in Psychology*, *61*, 100837. https://doi.org/10.1016/J.NEWIDEAPSYCH.2020.100837

Mashour, G. A., Roelfsema, P., Changeux, J. P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Mazor, M., Dijkstra, N., & Fleming, S. M. (2022). Dissociating the Neural Correlates of Subjective Visibility from Those of Decision Confidence. *Journal of Neuroscience*, *42*(12), 2562–2569. https://doi.org/10.1523/JNEUROSCI.1220-21.2022

Mazor, M., Friston, K., & Fleming, S. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. *eLife*, *9*, 1–34. https://doi.org/10.1101/853366

Merten, K., & Nieder, A. (2012). Active encoding of decisions about stimulus absence in primate prefrontal cortex neurons. *Proceedings of the National Academy of Sciences*, *109*(16), 6289–6294. https://doi.org/10.1073/pnas.1121084109

Meyen, S., Zerweck, I. A., Amado, C., von Luxburg, U., & Franz, V. H. (2022). Advancing research on unconscious priming: When can scientists claim an indirect task advantage? *Journal of Experimental Psychology. General*, *151*(1), 65–81. https://doi.org/10.1037/xge0001065

Morales, J. (2022). Introspection Is Signal Detection. *British Journal for the Philosophy of Science*. https://doi.org/10.1086/715184

817 Morales, J., Lau, H., & Fleming, S. M. (2018). Domain-general and domain-specific patterns of activity
818 supporting metacognition in human prefrontal cortex. *Journal of Neuroscience*, *38*(14),
819 3534–3546. https://doi.org/10.1523/JNEUROSCI.2360-17.2018
820 Park, S. A., Miller, D. S., Nili, H., Ranganath, C., & Boorman, E. D. (2020). Map Making: Constructing,
821 Combining, and Inferring on Abstract Cognitive Maps. *Neuron*, *107*(6).
822 https://doi.org/10.1016/j.neuron.2020.06.030
823 Persaud, N., Davidson, M., Maniscalco, B., Mobbs, D., Passingham, R. E., Cowey, A., & Lau, H. (2011).
824 Awareness-related activity in prefrontal and parietal cortices in blindsight reflects more than
825 superior visual performance. *NeuroImage*, *58*(2).
826 https://doi.org/10.1016/j.neuroimage.2011.06.081
827 Peters, M. A. K., & Lau, H. (2015). Human observers have optimal introspective access to perceptual
828 processes even for visually masked stimuli. *eLife*, *4*(OCTOBER2015).
829 https://doi.org/10.7554/eLife.09651
830 Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*.
831 https://doi.org/10.1146/annurev-neuro-071013-014030
832 Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human Orbitofrontal Cortex Represents a
833 Cognitive Map of State Space. *Neuron*, *91*(6). https://doi.org/10.1016/j.neuron.2016.08.019
834 Shekhar, M., & Rahnev, D. (2018). Distinguishing the roles of dorsolateral and anterior PFC in visual
835 metacognition. *Journal of Neuroscience*, *38*(22). https://doi.org/10.1523/JNEUROSCI.3484-
836 17.2018
837 Sladky, R., Friston, K. J., Tröstl, J., Cunnington, R., Moser, E., & Windischberger, C. (2011). Slice-timing
838 effects and their correction in functional MRI. *NeuroImage*, *58*(2), 588–594.
839 https://doi.org/10.1016/J.NEUROIMAGE.2011.06.078
840 Taschereau-Dumouchel, V., Cortese, A., Lau, H., & Kawato, M. (2021). Conducting decoded
841 neurofeedback studies. *Social Cognitive and Affective Neuroscience*, *16*(8).
842 https://doi.org/10.1093/scan/nsaa063

Tsuchiya, N., Wilke, M., Frässle, S., & Lamme, V. A. F. (2015). No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends in Cognitive Sciences*, *19*(12), 757–770. https://doi.org/10.1016/j.tics.2015.10.002

van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., Dehaene, S., & Roelfsema, P. R. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, eaar7186. https://doi.org/10.1126/science.aar7186

Weiskrantz, L., Warrington, E. K., Sanders, M. D., & Marshall, J. (1974). Visual capacity in the hemianopic field following a restricted occipital ablation. *Brain*, *97*(1), 709–728. https://doi.org/10.1093/brain/97.1.709

Whyte, C. J., & Smith, R. (2021). The predictive global neuronal workspace: A formal active inference model of visual consciousness. *Progress in Neurobiology*, *199*, 101918. https://doi.org/10.1016/j.pneurobio.2020.101918

Conflict of interest

The authors declare no conflict of interest.

Acknowledgments

ND is supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 882832. PK is supported by a Wellcome/Royal Society Sir Henry Dale Fellowship (218535/Z/19/Z) and a European Research Council (ERC) Starting Grant (948548). SMF is a CIFAR Fellow in the Brain, Mind and Consciousness Program, and supported by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [selected as ERC Consolidator, grant number 101043666]. The Wellcome Centre for Human Neuroimaging is supported by core funding from the Wellcome Trust (203147/Z/16/Z). The Max Planck UCL Centre is a joint initiative supported by UCL and the Max Planck Society. For the purposes of Open Access, the author has applied a CC-BY public copyright license to any author accepted manuscript version arising from this submission.

Appendix A

Table 1. Stimulus-encoding RSA clusters. All clusters are significant at $p < 0.05$ FWE corrected for multiple comparisons within the whole brain volume, with a cluster-forming threshold of $p < 0.001$ uncorrected. Clusters are included in the table if they surpass a threshold of 50 voxels. T-values, labels and coordinates are given for the peak within each cluster.

N voxels	t-value	AAL label	X Y Z
7541	9	Cuneus R	8 -74 40
6226	7.8	Hippocampus R	20 -39 1
5986	7.8	Cerebellum L	-1 -79 -8
5921	8.2	Cerebellum Crust1 R	38 -74 -22
4289	8.6	Hippocampus L	-23 -38 4
3068	6.9	Cerebellum L	-32 -64 -24
1129	7.4	-	-4 19 37
1121	7	Vermis 8	-5 -63 x -31
877	7	-	2 9 -5
812	7.2	Cingulum Mid R	3 -16 34
612	6.8	Frontal Mid L	-43 34 32
560	6.7	Cerebellum L	-32 -43 -27
498	7.1	Precuneus L	-10 -43 -27
372	7.8	Parietal Sub L	-22 -64 51
347	7.8	-	-23 22 -5
126	6.4	-	-15 -2 25
107	7	Cerebellum R	25 -53 -41
104	6.4	Cerebellum R	-14 -64 -36
93	6.7	Frontal Inf Oper L	-43 9 23
79	6.3	Cerebellum L	-12 -56 -48
55	6.6	Frontal Inf L	-43 23 22

Table 2. Detection prior RSA clusters. All clusters are significant at $p < 0.05$ FWE corrected for multiple comparisons within the whole brain volume, with a cluster-forming threshold of $p < 0.001$ uncorrected. T-values, labels and coordinates are given for the peak within each cluster.

N voxels	t-value	AAL label	X Y Z
318	7.2	Rectus R	10 32 -18
119	6.3	Frontal Mid Orb R	28 39 -12

885 **Table 3. Content prediction error RSA clusters.** All clusters are significant at p < 0.05 FWE corrected for multiple
886 comparisons within the whole brain volume, with a cluster-forming threshold of p < 0.001 uncorrected. Clusters
887 are included in the table if they surpass a threshold of 50 voxels. T-values, labels and coordinates are given for
888 the peak within each cluster.

N voxels	t-value	AAL label	X Y Z
10443	10.1	Precuneus	8 -72 46
8147	8.2	Putamen	22 -40 -2
4441	7.9	Cerebellum Crus R	39 -74 -22
3506	8.7	Cingulum Mid L	1 18 68
1951	9.4	Parietal Sup L	-20 -69 50
1899	8.0	Cingulum Mid R	2 -17 34
1370	7.4	Parietal Sup R	31 -55 57
902	6.8	-	-17 18 -5
368	6.7	Occipital Inf L	-43 -75 -10
358	6.4	-	-14 -2 24
299	6.5	Frontal Sup R	22 54 32
206	6.3	Frontal Mid L	-41 33 33
144	6.3	Frontal Sup Medial L	-1 60 15
140	6.7	Precuneus L	-10 44 72
122	6.5	Frontal Mid L	-26 40 30
115	6.7	Frontal Inf Oper L	-42 8 24
65	6.2	Caudate R	-7 13 8

889

890 **Table 4. Detection prediction error RSA clusters.** All clusters are significant at p < 0.05 FWE corrected for
891 multiple comparisons within the whole brain volume, with a cluster-forming threshold of p < 0.001 uncorrected.
892 Clusters are included in the table if they surpass a threshold of 50 voxels. T-values, labels and coordinates are
893 given for the peak within each cluster.

N voxels	t-value	AAL label	X Y Z
6599	8.9	Rectus R	2 40 15
4616	8.7	Cerebellum	-31 -51 24
2488	7.7	Angular L	-39 -69 45
1757	7.0	Cerebellum R	21 -44 -19
1340	6.7	Precuneus R	4 -56 21
1252	7.4	Frontal Mid Orb R	33 40 -11
1077	7.2	Frontal Mid L	-39 13 34
609	7.1	Vermis	-3 -64 -31

302	6.6	Amygdala R	28 1 -14
269	6.7	Frontal Sub Orb L	-33 59 -2
223	7.9	-	-23 -18 20
221	6.8	-	16 -2 25
145	6.3	Caudate L	13 15 10
139	6.9	Precuneus L	-12 -70 39
120	6.5	Frontal Mid L	-26 46 39
99	6.5	Frontal Mid R	32 50 12
97	6.5	Cerebellum L	-14 -53 -14
89	6.6	Cerebellum R	26 -50 -42
86	6.7	-	-16 -36 42
67	6.9	Hippocampus L	-35 -35 -6

894

895

896