Representation of sustained visual experience by time-invariant distributed neural patterns

Gal Vishne[1,*], Edden M. Gerber[1], Robert T. Knight[2,3], and Leon Y. Deouell[1,4,*]

[1] Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel
[2] Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, USA
[3] Department of Psychology, University of California, Berkeley, Berkeley, CA, USA
[4] Department of Psychology, The Hebrew University of Jerusalem, Jerusalem, Israel
* Corresponding author: gal.vishne@mail.huji.ac.il, leon.deouell@mail.huji.ac.il

Abstract

Everyday visual experience contains many brief stationary moments when we continue to be consciously aware our sensory input. However, neural activity is dramatically attenuated shortly after the onset response, and it is unclear how the content of our experience remains stable nevertheless. To characterize the spatiotemporal dynamics of object representation in sustained viewing conditions, we use intracranial recordings from ten patients viewing images from variable categories in multiple durations. We reveal that the distributed neural representation of categories and exemplars in posterior-sensory regions remains sustained and stable, mirroring the stability of experience, despite substantial variability in activity after the onset response. Additionally, we find transient coding of visual content in frontoparietal cortex, even when no report is required. Together, our results suggest sustained perception is maintained by sensory regions, not by overall activation magnitude but by the position within an "Experience Subspace", and perceptual changes involve frontoparietal regions, regardless of the task. These findings directly inform recent large-scale attempts to test theories of consciousness.

Introduction

Visual recognition is a remarkably rapid and efficient process. We recognize images flashed for less than 20ms, and signatures of recognition in the brain evolve as early as 150ms after the onset of a visual stimulus[1,2]. Yet, subjective visual experience is more than a series of flashing or ever- changing objects. Rather, it continues uninterrupted between moments of change. We often fixate for several seconds on the same static object or scenary[3,4], and continue to experience it. Even if we consider saccades as triggering a change in perception, these occur only 2-4 times a second, leaving the majority of perceptual time to a mostly static percept. Despite this state of affairs, most of what we know about visual neuroscience pertains to onset or change-related activity[5–7]. What happens beyond the onset, as we continue to perceive the world? The answer to the question is especially intriguing considering that activity in category selective regions of in the ventral temporal cortex (VTC), presumed to be integral to our ability to recognize objects, goes down to almost floor levels after the initial onset response, independent of the duration of the percept[8,9]. Given this dissociation

between experience and neural activity, it is still unknown how the content of visual experience is represented during these frequent periods of sustained visual input.

This question has important implications for the study of the "neural correlates of consciousness" (NCC), which are the minimal set of neuronal mechanisms jointly sufficient for any one specific percept[10–12]. Several candidate theories have been proposed, but the field is far from convergence[13]. One prevalent approach ties conscious perception to a non-linear "ignition" of neural activity[14], characterized by strong recurrent connections, either within sensory areas[15,16], or in a broad network of executive areas, known as the Global Neuronal Workspace (GNW)[17,18]. A major challenge to this approach is that the temporal dynamics of ignition do not correspond to the temporal dynamics of conscious percepts[19], as neural ignition is typically a few hundred milliseconds long, regardless of whether the stimulus is experienced for a few tens of milliseconds[16], or for longer than a second[8,9]. A competing major theory, the Integrated Information Theory (IIT)[20] contends that the NCC relies on a cause-effect structure of integrated information which should hold as long as the experience does[12], but the evidence is still lacking. Another major point of disagreement is regarding the anatomical basis of the NCC. Some theories, most notably GNW theory[17,18] and higher-order theories (HOTs)[21,22], emphasize the crucial role of the prefrontal cortex (PFC). Other theories (e.g., IIT[20] and recurrent processing theory[15]) maintain in contrast that prefrontal activation is not essential for conscious awareness[23], and instead emphasize activity in sensory areas, or a posterior "hot-zone", as the proper basis of phenomenal experience, based on its structure[24]. By this account, the observed prefrontal activations[25–27] reflect mainly the need to report on a percept, and are reduced or eliminated when such report is not required[28,29]. We conjectured that whatever account of NCC is right, it should account for the persistent experience and not only to experiencing onsets or changes. Moreover, a full isomorphism between experience and neural representation requires that the representation will remain constant, mirroring the unchanging content of perception. We therefore designed our experiment to examine the response to highly visible images of different durations and its anatomical basis[9]. Our experimental logic was recently adopted by the COGITATE consortium's adversarial collaboration, which also provided explicit predictions about the outcome by proponents of the GNW and IIT[12,30].

We used electrocorticography (ECoG) recordings from ten patients undergoing surgery for intractable epilepsy[31] (>1000 electrodes, distributed across all cortical lobes; Figure 1a), while they were viewing a diverse set of images belonging to multiple visual categories (Figure 1b), and presented for various durations (300-1500ms, Figure 1c). The diversity of images allows us to track visual representation at multiple levels of abstraction, from superordinate categories (e.g., animate\inanimate) to basic categories (e.g., faces\animals) to single exemplars (e.g., specific faces). The diversity of presentation durations allows us to address the correspondence of experience with neural activity, distinguishing between responses to the onset of a new stimulus and neural signals corresponding to the dynamics of ongoing perception. We avoid the conundrum of conflation with report-related activity by using above-threshold, clearly visible stimuli, which permits us to forgo trial-by-trial awareness probes and analyze only trials not requiring responses. In our initial report of this data, we confirmed the drastic reduction of broadband high frequency activity following the onset response in higher order, category selective visual cortex[9] (see [32] for an fMRI follow up). Here, we moved to examine multivariate representations in a larger set of regions of interest.

We find that sustained visual content is represented in posterior-sensory regions by the pattern of responses across electrodes, which enables a temporally stable readout of both category and exemplar-level information corresponding to the ongoing percept, despite considerable attenuation in activity magnitude occurring simultaneously. This suggests that posterior-sensory regions maintain our ongoing perceptual experience using a distributed code, independent of the overall response magnitude. In addition to the posterior representation, we find transient representation of visual content in several cortical areas, including PFC, despite the lack of report, which suggests involvement of these regions in experiences of perceptual changes, regardless of the task. Together, our findings confirm predictions of both major theories of consciousness (IIT and GNWT) provided in the recent adversarial collaboration[12,30], suggesting that these predictions alone cannot arbitrate between the theories. Rather, it is possible that transient and sustained representations, in posterior and prefrontal areas respectively, support different aspects of conscious awareness. By studying visual experience beyond the onset, we provide a critical building block to the study of the neural correlates of consciousness.

Results

Visually responsive electrodes are present in all cortical lobes

We recorded local field potentials from ten patients with drug-resistant epilepsy implanted with subdural ECoG electrodes for clinical purposes (Supplementary Table 1). Across patients, electrodes covered all cortical lobes (Figure 1a; see Supplementary Tables 2-3 for electrode distribution per region and per patient). Patients viewed grayscale images of variable durations (300 to 1500ms) belonging to multiple visual categories including faces, man-made objects, and animals (Figure 1b; Supplementary Table 4). Within the object category, we analyzed separately a sub-category of watches, which bear low level visual similarity to faces[33,34]. Responses were required for less than 10% of the trials: appearance of a rare category (clothing items, all patients), or blurring near the end of the presentation (seven of the patients; see Methods and Figure 1c). Detection of clothing items was high (mean ± SEM across patients: 87.8 ± 1.3%), and false alarm rates were low (3.2 ± 0.3%) in all patients. Targets of both types were excluded from further analysis, focusing on images which did not require any report. For all analyses, we utilize high-frequency activity (HFA, 70-150Hz), shown to track neuronal firing rate in humans[35–37] (see ref [38] for more refined laminar distinctions).

Of 907 noise-free electrodes, we focus on the 430 responsive electrodes, liberally defined as significant HFA modulation relative to a 200ms pre-stimulus baseline, for at least one of the categories, in at least one of four non-overlapping 200ms time-windows between 100-900ms after stimulus onset (for stimuli presented for 900ms or longer, *two-tailed* paired t-test; corrected for multiple comparisons across electrodes (FDR)[39] and windows (Bonferroni); for the distribution of responsive electrodes to each category in each time-window see Supplementary Table 5). Thus, the selection of electrodes for analysis did not include any category or temporal selectivity measures. Responsive electrodes were found in all patients and in all lobes where activity was recorded (Figure 1a), with occipital (Occ) and ventral-temporal cortex (VT) containing the largest percentages of responsive electrodes, as expected (79% and 71% of all non-noisy electrodes in the region, respectively), followed by prefrontal (PFC, 36%), parietal (Par, 31%), lateral-temporal (LT, 20%), and sensorimotor cortex (SM, 15%).

We begin by examining the single electrode response dynamics (which we previously explored in [9]), followed by analyses of the multivariate responses across electrodes. Activity in 80% (346/430) of sites increased relative to the pre-stimulus state, therefore our analysis of single-electrode response dynamics focuses on this electrode group in the four regions showing substantial visual activity (Occ, VT, PFC and Par; for results in SM and LT see Supplementary Figure 1a-d). For sites where activity significantly decreased relative to pre-stimulus state see Supplementary Figure 1e. All responsive electrodes (increasing or decreasing activity) were included in the single-electrode category selectivity analysis and all multivariate analyses (control analyses indicated this decision did not affect conclusions).

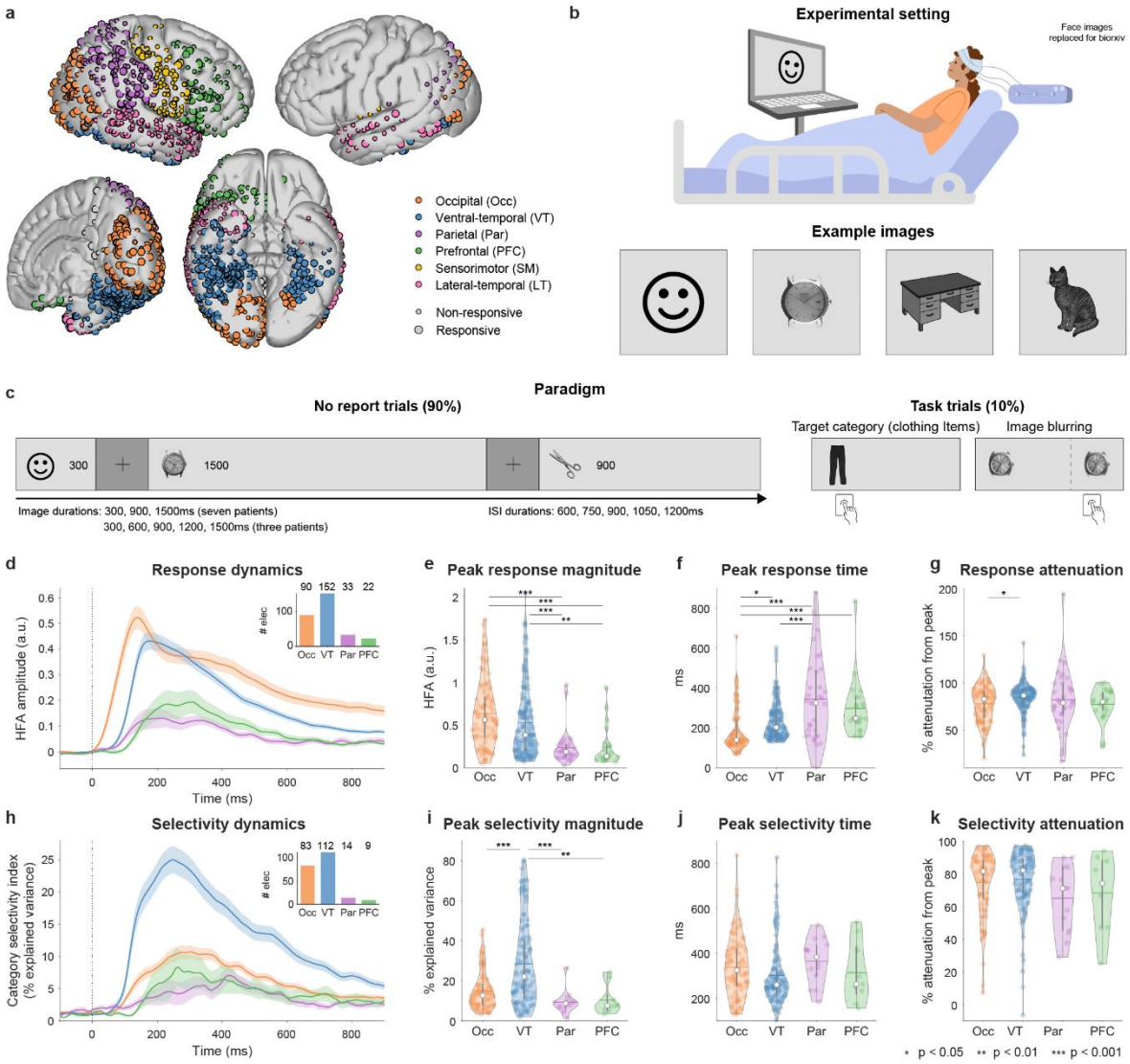

Figure 1. Experimental setup and design, electrode locations and single electrode (univariate) properties. a Electrode locations (pooled across patients). Electrodes are colored according to their anatomical location (Occ, occipital; VT, ventral-temporal; Par, parietal; PFC, prefrontal; SM, sensorimotor; LT, lateral-temporal). Large markers correspond to responsive electrodes (see Supplementary Table 2 for the number of electrodes in each ROI). **b** Experimental setup and design. Top: recording was conducted in the epilepsy ICU, with stimuli presented on a laptop screen. Bottom: example images from the four main categories (faces, watches, objects and animals). **c** Left: illustration of three passive (no report) trials. Images were presented for variable durations

(300, 900, 1500ms and for three patients also 600, 1200ms), with variable inter-stimulus-intervals (ISI; 600, 750, 900, 1050, 1200ms). Right: illustration of the two target types requiring a button press (together 10% of trials): 1. Clothing items 2. Blurring of the image. **d-g** HFA response magnitude (relative to pre-stimulus baseline; only responses to images with duration of at least 900ms). Only electrodes showing an increase in response relative to the pre-stimulus period were included in the calculation (number in each region shown in the inset of **d**), and for each electrode we considered only trials from categories the electrode was responsive to. **d** HFA activation dynamics. Time zero denotes stimulus onset. Single trial responses were averaged for each electrode, shaded area denotes SEM across electrodes. **e** Peak HFA response magnitude. **f** Time of peak HFA response. **g** Relative attenuation in HFA responses from peak to the end of stimulus presentation, higher numbers indicate stronger attenuation (peak activity minus mean activity 800-900ms after onset relative to peak activity; segments were baseline corrected, attenuation is >100% when the activity 800-900ms is lower than the baseline levels). **h-k** Category selectivity (only responses to images with duration of at least 900ms). Only category selective electrodes were included in the calculation (as the goal was to estimate the dynamics of selective electrodes, not the absolute selectivity; number in each region shown in the inset of **h**). **h** Category selectivity over time (Eta-squared from a time-point by time-point ANOVA, expressed as percent explained variance). **i** Peak selectivity magnitude. **j** Time of peak selectivity. **k** Relative attenuation in selectivity, higher numbers indicate stronger attenuation (peak selectivity minus mean selectivity 800-900ms after onset, relative to peak selectivity). Violin plots (**e-g**, **i-k**) were created using 'violinplot.m' in Matlab[40] – each colored dot stands for a single electrode, horizontal lines of the same color denote the mean across electrodes, white dots denote the median and gray vertical bars denote the interquartile range. Contour lines denote kernel probability density estimate (Matlab function ksdensity). Regions connected by a black horizontal line showed significant post-hoc differences (Tukey-Kramer method).

Response magnitude and category selectivity in single electrodes is substantially attenuated after the initial onset response, despite persistent visibility of the image

Figure 1d depicts the HFA response dynamics in different regions, for stimuli presented for 900ms or longer (see also Supplementary Figure 1a, e-f). As expected, response magnitude was substantially higher in the classical visual regions VT and occipital cortex (peak HFA amplitude, mean ± SEM across electrodes: Occ, 0.63 ± 0.04; VT, 0.52 ± 0.03, arbitrary units (a. u.) due to the normalization procedure), relative to parietal and prefrontal cortex (Par: 0.23 ± 0.03; PFC: 0.24 ± 0.05), and this difference was highly significant (one-way ANOVA comparing the four regions depicted in Figure 1e, $F(3,293) = 12.5$, $p < 10^{-6}$), with post-hoc tests (Tukey-Kramer method, throughout the paper) showing significant differences between the visual regions and the frontoparietal responses (all $p < 0.006$, all Cohen's $d > 0.7$), but not between VT and Occ or PFC and Par. Responses peaked first in Occ (mean ± SEM across electrodes: 187 ± 12ms; Figure 1e), followed by VT (233 ± 8ms), PFC (297 ± 31ms) and Par (343 ± 36ms). These differences were significant (Figure 1f; one-way ANOVA, $F(3,293) = 15.73$, $p < 10^{-8}$), and post-hoc tests showed significant differences between occipital cortex and all other regions ($p_{Occ\text{-}VT} < 0.019$, Cohen's $d = 0.47$; $p_{Occ\text{-}Par} < 10^{-9}$, $d = 1.1$; $p_{Occ\text{-}PFC} < 0.001$, $d = 0.95$), and between VT and parietal cortex ($p_{VT\text{-}Par} < 0.0001$, $d = 0.9$). Similar patterns were observed for half-peak times which are less sensitive to residual noise (Supplementary Figure 1g). Relative to the peak response, mean response magnitude 800-900ms after onset was attenuated by 82.4 ± 1.1% (mean ± SEM of electrodes from all six ROIs), despite the continued visibility of the stimulus on the screen. The level of attenuation varied across regions (one-way ANOVA, $F(3,293) = 2.98$, $p < 0.033$; Figure 1g). The signal was most attenuated in VT (84.7 ± 1.2%), followed by Par (82 ± 5.7%), Occ (77.7 ± 2%) and PFC (77.3 ± 3.8%), yet post-hoc tests revealed only a difference between VT and Occ ($p < 0.03$, Cohen's $d = 0.42$). Thus, activity drastically declined after the initial onset response in all regions.

A reduction in response magnitude over time does not necessarily imply a reduction in category information over time. If variability within categories is attenuated alongside the mean response, then the overall

discriminability between categories will remain high. To test for this, we computed the category selectivity of each responsive electrode (including both positively and negatively responding sites) in four non-overlapping 200ms time-windows between 100-900ms after onset (similarly to the responsiveness definition). In each time-window we computed a one-way analysis of variance comparing the mean responses to each category. Of the 430 responsive electrodes, 236 electrodes (55%) showed significant category selectivity in at least one of the time-windows ($q_{FDR} < 0.05/4$). As expected, most category selective electrodes were in VT (112, 47% of all category selective electrodes), followed by occipital cortex (83, 35%), and with much lower frequency, the other ROIs (PFC, 9 electrodes; Par, 14; SM, 2; LT, 10; Medial, 6; see Supplementary Figure 1h for more details). We computed a selectivity time-course for each category selective electrode using Eta-squared (η^2) from a time-point by time-point ANOVA, reflecting the proportion of between trial variance explained by the visual category (Figure 1h). Peak selectivity differed between regions (one-way ANOVA, $F(3,214) = 14.84$, $p < 10^{-8}$; Figure 1i). Post-hoc tests showed higher selectivity in VT (% of variance explained by visual category, mean ± SEM across electrodes: 28.5 ± 2%) relative to all other regions (Occ, 15 ± 1.1%; Par, 9.3 ± 1.5%; PFC, 10.3 ± 2.5%; $p_{VT\text{-}Occ} < 10^{-6}$, $d = 0.78$; $p_{VT\text{-}Par} < 0.0003$, $d = 0.95$; $p_{VT\text{-}PFC} < 0.009$, $d = 0.88$) with no other significant pairwise comparison. Category selectivity peaked later than raw HFA response, roughly at the same time across the cortex, (Figure 1j; peak mean ± SEM across electrodes from all ROIs: 323 ± 9ms; no significant differences between ROIs, $F(3,214) = 1.82$, $p > 0.14$; half-peak times significantly earlier in VT relative to Par, $F(3,214)=4.82$, $p < 0.003$; $p_{VT\text{-}Par} < 0.008$, $d = 0.92$, no other significant comparisons). In line with the general reduction in the magnitude of responses, single electrode category selectivity also declined drastically, by an average (± SEM) of 74.4 ± 1.3% from the peak selectivity time to 800-900ms after stimulus onset (Figure 1k), with no significant differences between regions (one-way ANOVA $F(3,214) = 1.89$, $p > 0.13$). Pairwise comparison of each category pair separately showed similar patterns (Supplementary Figure 1i-j). To conclude, both response amplitude and category selectivity in single electrodes were substantially attenuated after the peak response, despite the continued presence of the stimulus on the screen.

Following the onset response, neural trajectories in posterior sensory regions settle into a sustained response state which tracks the duration of the stimulus

To understand how information is encoded in patterns of distributed activity, we first examined the multivariate state-space trajectory[41–43] in response to each category separately, for stimuli lasting 900ms or longer (Figure 2a-c, for regions SM and LT see Supplementary Figure 2a-c). We quantified the global multivariate response profile by examining the time-point by time-point distance of the neural state from the neural baseline state (when no visual stimulus was presented; Figure 2b). Multivariate responses peaked earlier than the univariate responses, though the relative order between regions was maintained, with responses peaking first in occipital cortex (131-150ms across categories), followed by VT (160-240ms) and parietal cortex (169-196ms), and finally PFC (234-275ms). Here too, we observed substantial attenuation after the initial onset response – relative to the peak response, the multivariate response after 800-900ms was attenuated by approximately 80% in all regions except Occ where it was attenuated by ~60% (Supplementary Figure 2d). Corresponding to the multivariate response dynamics, state transitions (from baseline to peak response) were fastest in occipital cortex (time of peak state transition speed across categories: 73-100ms),

followed by VT (117-120ms) and parietal cortex (113-130ms) and slowest in PFC (133-152ms; Figure 2c; see [44] for a similar approach). Returning to the pre-stimulus state was slower, especially in VT and Occ (trajectory speed up to peak response was significantly higher than trajectory speed after peak response for all regions and categories except for two cases; Supplementary Figure 2e).

The slow rate of change after the initial onset response, as well as examination of the multivariate trajectories, suggest that the response trajectories in some regions settle into a relatively sustained activity mode, which does not converge to the pre-stimulus state while the image is still presented. To further explore this possibility, we examined the response trajectories to each stimulus duration separately (300, 900 and 1500ms), focusing here on the faces category (similar results were obtained for other categories, Supplementary Figure 3c-d). If the representative neural pattern persists throughout stimulus presentation, and subsides shortly after its offset, the global multivariate response (distance from the pre-stimulus state) is expected to be larger when a stimulus is currently displayed than when it has already disappeared. Thus, we examined the *difference* in global multivariate response between the trajectories for stimuli that are 900ms and 300ms in duration, and for stimuli that are 1500ms and 900ms long. Both comparisons were significant in ventral-temporal and occipital cortex, with significance starting shortly after the short stimulus was taken off the screen and persisting throughout the entire time the long stimulus was presented (Figure 2d; time-resolved permutation test with max-statistic control for multiple comparisons, $p < 0.05$). Other regions did not show such duration sensitivity (Figure 2d and Supplementary Figure 3a). Additionally, we computed the time-point by time-point distances between the trajectories of different durations (without considering the distance to the pre-stimulus state), which supported the same conclusion (Supplementary Figure 3b, d). Together, these results suggest that when a stimulus is displayed for prolonged time-periods, posterior sensory regions (VT and Occ) track the stimulus duration by transitioning to a sustained multivariate response state characterized by a constant distance from the pre-stimulus state and little to no state changes, and returning to the pre-stimulus state only after the stimulus is removed from the screen. In contrast, neural trajectories in PFC and other regions do not show dependence on the duration of the stimulus, suggesting that these areas only reflect perceptual changes.

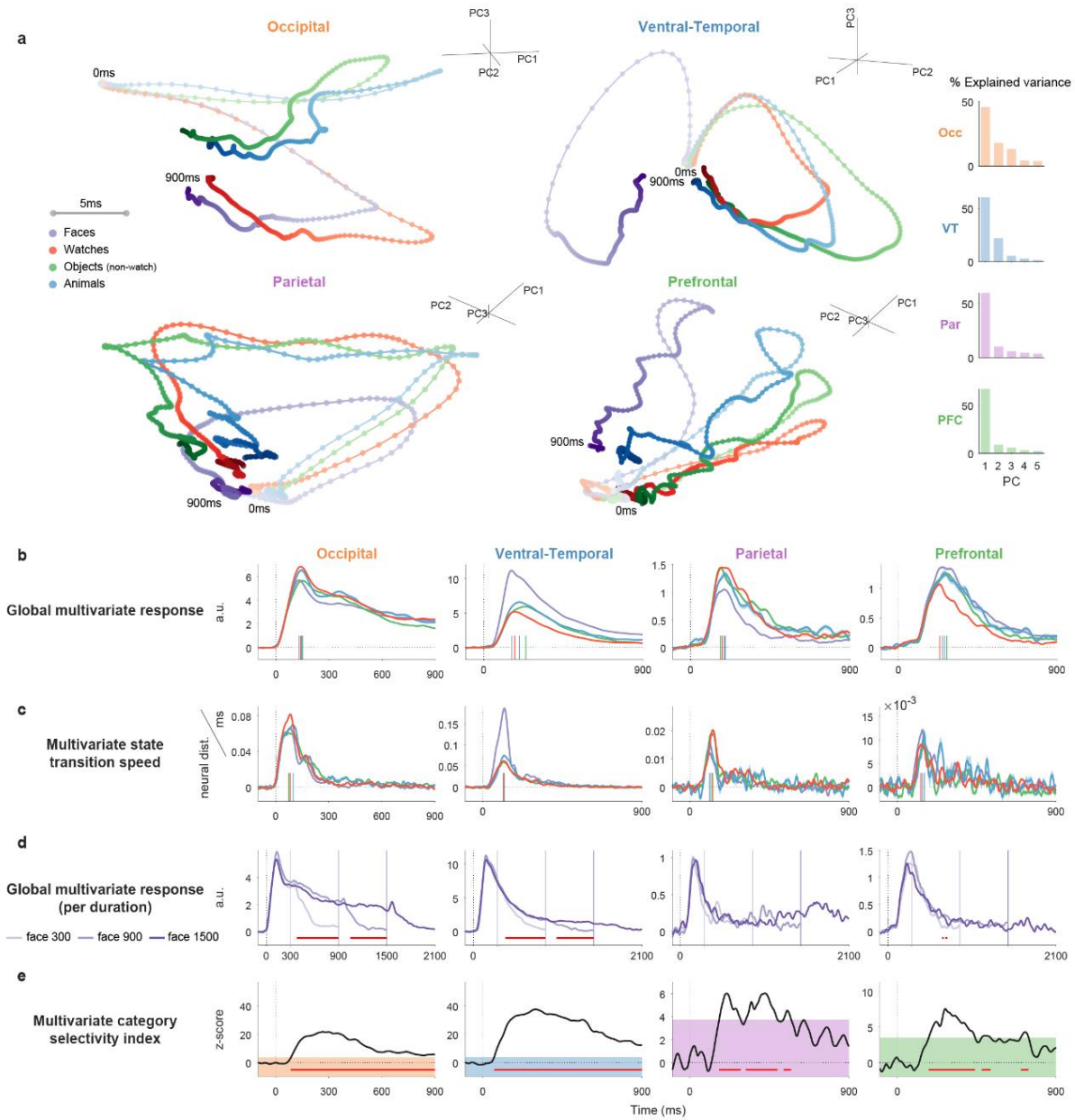

Figure 2. Multivariate state-space dynamics show sustained responses in posterior regions and distributed coding of category information, including PFC. a Multivariate response trajectories (projection on first 3 principle components) to each of the four image categories, for stimuli presented for 900ms or longer (for corresponding plots of the other regions see Supplementary Figures 2-3). Each line corresponds to a different category. Time is marked by the color and width of the trajectory (darker and thicker as the trial progresses). Dots are separated in time by 5ms. Larger separation between dots indicates faster speed of change in state-space. PCA was performed in each region separately using all responsive electrodes and was used for visualization purposes only; characterization of response dynamics (**b-e**) was carried out using the full multivariate response. Responses to each category were averaged prior to the calculation. Viewing angle in each region was chosen to best visualize differences between categories. Bar graphs on the right depict the explained variance of the first five components for each region. **b-e** Characterization of the multivariate response dynamics. Note that amplitudes are not comparable between regions, only within the same region in different time-points, as these metrics are also dependent on the number of electrodes in each region. **b** Global multivariate response dynamics per-category (time-point by time-point distance of the state-space trajectory from the pre-stimulus state) for the trajectories depicted in **a**, with the same color code. Colored vertical lines on the abscissa indicate peak times. Stimulus onset is marked by a dashed line. **c** State

transition speed dynamics ("neural distance" travelled by the trajectory per ms) for the trajectories depicted in **a** (notations as in **b**). State transition speed from the baseline state to the onset state was rapid, while return to baseline was more gradual. **d** Global multivariate response dynamics to face images per stimulus duration (for other categories see Supplementary Figure 3). Darker lines correspond to longer stimuli. Trajectories are cropped 600ms after stimulus offset since this was the shortest ISI and beyond that point another stimulus could be presented. Stimuli offsets at 300, 900 and 1500ms are marked by a vertical line of the same color. Stimulus onset is marked by a dashed line. Horizontal red bars mark time points with significant differences between trajectories of different durations using permutation max-statistic control for multiple comparisons ($p < 0.05$). Only time-points after the offset of the shorter stimuli were considered in each comparison, therefore points in the interval 300-900ms mark significant differences between 300 and 900ms trajectories, and points in the interval 900-1500ms mark significant differences between 900 and 1500ms trajectories (results are highly similar considering the entire presented range). **e** Category selectivity index (measuring the dispersion between responses to different categories) was significantly above chance in all regions. Category selectivity traces were z-scored on a time-point by time-point basis using the null distribution obtained from shuffling the category labels (similar results were obtained without this step). Only stimuli with duration ≥ 900ms were used in the calculation (as in **a-c**). Horizontal red bars indicate above chance category selectivity index, using permutation max-statistic control for multiple comparisons ($p < 0.05$). The shaded regions mark the 95% percentile of the null distribution. **b-d** Traces were baselined for presentation purposes.

Visual category information is encoded in a distributed fashion across the cortex

The previous section established the temporal dynamics of the multivariate neural state in different regions. Next, we asked whether, where, and when these neural states distinguish between categories. If a region contains information about the visual category of the image, the state-space trajectories in response to various categories should be distinguishable. To test this, we examined the time-point by time-point dispersion between the trajectories of the different categories (Category selectivity index; Figure 2e and Supplementary Figure 3e; stimuli with duration ≥ 900ms). We found significant between-category variability in posterior sensory regions, which persisted throughout the presentation of the stimulus (Occ, 87-900ms; VT, 65-900ms; permutation test with max-statistic control for multiple comparisons, $p < 0.05$), alongside more transient category information in frontoparietal regions (Par, 167-288ms, 317-496ms and 530-573ms; PFC, 178-440ms, 479-528ms and 701-744ms). The categorical representational structure varied between regions (Figure 2a). This can be gleaned, for example, from the fact that in VT the strongest category distinction between trajectories was between faces and all other categories, while in occipital cortex the trajectories of faces and watch stimuli grouped together, possibly due to the low-level similarities between the two categories. Prefrontal and parietal cortex both showed more distinct neural trajectories in response to face images compared with the other categories, but the underlying responses differed – in PFC the face category response diverged from the other categories immediately after stimulus onset, while in parietal cortex activity in response to faces resembled the trajectory of other categories but remained in a lower magnitude. This contrast stems, at least partially, from the fact that 36% of responsive electrodes in PFC responded only to faces, compared to just 9% in parietal cortex.

These results reveal that the average neural state-space trajectory varies depending on the viewed category and this information is present in multiple regions, including the parietal and prefrontal cortices. But is this information enough to enable trial-by-trial classification? To answer this question we used multivariate decoding on a subset of 92 unique images which were shown to all patients for 900ms or longer (Figure 3a; faces, 30; watches, 32; objects, 18; animals, 12)[45,46]. For each cortical region separately, we trained a set of linear classifiers to distinguish between each pair of categories on a time-point by time-point basis, using the

HFA responses across electrodes in each region as features. Classifier performance was evaluated using area under the curve (AUC) together with permutation testing (one-sided throughout the paper). Peak single trial decoding performance (Figure 3b) was high both in posterior sensory regions (peak AUC mean ± SEM across comparisons: VT, 99.6 ± 0.3%; Occ, 98.1 ± 1.6%; $p_{perm} < 0.002$ for all comparisons, max-statistic permutation test) and in frontoparietal regions (PFC, 84.1 ± 2.3%; Par, 86 ± 1.7%; $p_{perm} < 0.041$ for all comparisons except object-animal decoding (both regions), and watch-object decoding in PFC). Thus, category information was not limited to traditional visual areas.

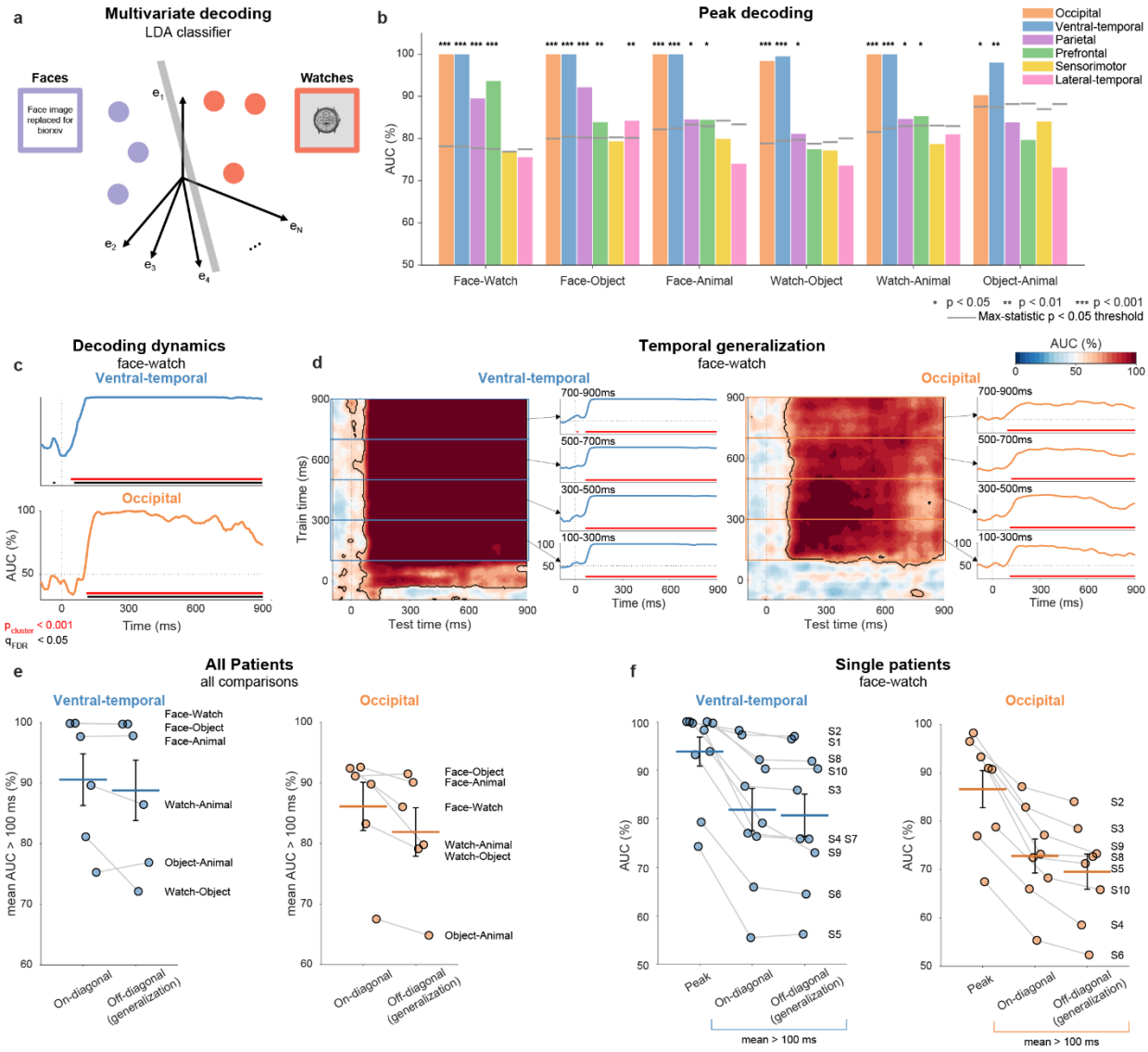

Figure 3. Category information is represented in multiple cortical regions; representation is sustained and stable in posterior sensory regions. a Schematic illustration of the analysis: We quantified category representation using time-point by time-point multivariate decoding. For each time-point we trained a linear classifier to identify the visual category of the presented image from the neural state space response (HFA activity across electrodes). Each colored dot represents the response in a single trial (color-coded according to the presented category), and the linear classifier is represented by a gray bar. Only images presented for 900ms or longer were included in analyses presented in this Figure. **b** Peak decoding performance for all pairwise comparisons in all ROIs showing significant category information in both posterior sensory regions and frontoparietal regions. Bars are color-coded

according to the region. Statistical significance assessed using a permutation test equivalent to a max-statistic control for multiple comparisons. Gray horizontal lines denote the threshold for significance (95% of permutation values). Thresholds are higher for comparisons involving categories with less exemplars. **c** Decoding dynamics (AUC time-course) in VT and Occ for the face-watch comparison reveal sustained coding of category information throughout stimulus presentation (for similar results with other pairwise comparisons see Supplementary Figure 4; see Figure 4a for PFC and Par). Dashed lines mark stimulus onset and chance level (AUC = 50%). Red horizontal bars mark significant clusters, black horizontal bars mark time-points significant with point-by-point permutation testing (FDR corrected). **d** Temporal generalization matrices in VT and Occ showing rectangular generalization profiles, indicating highly stable coding of category information (see Figure 4b for PFC and Par). Classifiers were trained for each time-points separately (marked by the y-axis position) and tested on all time-points (x-axis position). The diagonal of these matrices (training and testing on the same time-point) corresponds to the AUC time-courses depicted in **c**. Significance was tested using permutation testing with FDR correction for multiple comparisons (contiguous points with $q_{FDR} < 0.05$ are surrounded by black contour lines). Insets on the right depict the mean generalization dynamics for training on time-points between 100-300ms, 300-500ms, 500-700ms and 700-900ms. Dashed lines mark stimulus onset and chance level. Red horizontal bars mark testing points which were significant for at least 50% of training points in the averaged range. **e** Decoding on-diagonal (training and testing on the same time point) and off-diagonal (generalization; both mean after 100ms) for all pairwise comparisons in VT and Occ. Results corresponding to each comparison are connected by a gray line. The mean across comparisons is denoted by a horizontal line, error bars mark ± SEM. Both regions show sustained high decoding which was highly stable between time-points. The full generalization matrices for all comparisons are shown in Supplementary Figure 5a-b. **f** Single patient decoding results (face-watch comparison) for all patients with responsive electrodes in VT (10/10 patients) and Occ (8/10 patients). Notations as in **e**. The locations of responsive electrodes in each patient, along with the full dynamics and temporal generalization matrices are shown in Supplementary Figure 5d-e.

Sustained and stable coding of category information in occipital and ventral-temporal regions although response magnitude is substantially attenuated

Next, we turn to examine the temporal dynamics of decoding performance, focusing first on posterior-sensory regions (Par and PFC are discussed in the following section). Figure 3c shows decoding of faces versus watches, the two largest categories (other category pairs shown in Supplementary Figure 4). Significant clusters emerged early after stimulus onset and persisted throughout stimulus presentation (correction for multiple comparisons using cluster-based permutations[47]; significant clusters: VT, 40-900ms, Occ, 110-900ms, both $p_{cluster} < 0.001$). Since cluster-based statistics are not sufficient to establish the exact latency or temporal extent of effects[48], we repeated the same analysis using point-by-point permutation testing with FDR correction, which also indicated that category information was sustained in both regions (Figure 3c, black horizontal bars). Importantly, there was barely any decrease in category information throughout this time (face-watch mean AUC after 100ms: VT, 99.8%, Occ, 92.4%), despite the substantial attenuation in response magnitude and in single electrode selectivity observed in both regions (Figure 1d, h). Other category comparisons showed similar profiles of high and sustained decoding, including comparisons that did not involve the face category (mean AUC after 100ms (mean ± SEM across all comparisons): VT, 91 ± 4%, Occ, 86 ± 4%; Figure 3e). These decoding analyses included both positive and negative responsive electrodes, but using only electrodes with positive responses resulted in highly similar time-courses (Supplementary Figure 4). Within occipital cortex, both retinotopic and non-retinotopic regions showed similar profiles of high and sustained decoding (Supplementary Figure 5c). These results did not stem from a single patient exhibiting high separability between the categories (Figure 3f and Supplementary Figure 5d-e). On the contrary, we observed high decoding performance in the majority of patients with electrodes in VT (peak AUC > 93% for 8/10 patients; mean ± SEM across patients: 93.8 ± 3%) and Occ (peak AUC > 90% for 5/8 patients; mean ± SEM across patients: 86.6 ± 3.9%). In most patients, high decoding persisted

throughout stimulus presentation with high AUC and significant clusters extending until 900ms (VT 7/10 patients; Occ 6/8 patients).

To test whether the sustained category discriminability we observed stems from a single sustained distinguishing pattern or from a sequence of changing states, we applied the temporal generalization method[49], which consists of two stages: (1) classifiers are trained on data from each time-point separately (as in the standard decoding analysis) and (2) each classifier is tested on *all* time-points. Successful decoding between time-points (off-diagonal decoding) indicates the pattern separating the two categories is stable in time. The rectangular temporal generalization pattern we found in both regions indicates that classifiers trained during the onset response were able to distinguish between the categories during the sustained response and vice versa, revealing a highly stable representation of category information (Figure 3d depicts the temporal generalization matrix for the face-watch comparison, other comparisons are shown in Supplementary Figure 5a-b; statistical significance in these panels was assessed using point by point permutation testing with FDR correction, $q_{FDR} < 0.05$). This pattern was especially striking in VT, where generalization between time-points was near perfect, and not significantly different from mean performance of classifiers trained and tested using data from the same time point (permutation test, $p_{perm} > 0.4$), for all face related comparisons (mean off-diagonal AUC (generalization) from 100ms and on ≥ 97.8% in all three cases; Figure 3e). Generalization was high for non-face related comparisons as well (mean off-diagonal AUC after 100ms (mean ± SEM across all comparisons): 88.8 ± 5%; Supplementary Figure 5a). Category information was also stable in occipital cortex (mean off-diagonal AUC after 100ms: 81.9 ± 4%; significant generalization for more than 85% of time-points after 100ms, except in the object-animal comparison; Figure 3e and Supplementary Figure 5b). We found similar results for retinotopic and non-retinotopic sub-regions of the occipital cortex (Supplementary Figure 5c). Results in both VT and Occ were highly consistent at the single patient level, where mean off-diagonal AUC after 100ms was as high as 97% in some single patients (Figure 3f; mean ± SEM across patients: VT 81.9 ± 4%, Occ 69.6 ± 4%; see Supplementary Figure 5d-e for the full generalization matrices of each patient).

Thus, although the magnitude of activity changed considerably in both regions (Figure 1h, Figure 2f), the pattern across electrodes remained consistent, enabling a steady and reliable readout of category information throughout stimulus presentation using the same multivariate classifier.

Visual category information is decodable from prefrontal cortex, despite the lack of report in the task

We observed high decoding performance in both PFC and parietal cortex that was not the result of a single patient exhibiting high decoding performance (Supplementary Figure 6a-b). Focusing on the face-watch comparison (Figure 4a; for other comparisons see Supplementary Figure 4c-d), we observed significant decoding in PFC in a cluster extending from 155 to 600ms and in parietal cortex from 170 to 605ms (both $p_{cluster} < 0.001$; similar time ranges were observed using point-by-point permutation tests with FDR correction, Figure 4a black horizontal bars). Peak decoding was reached after 325ms in PFC, and 305ms in parietal cortex, both within the range of the proposed time for global neural ignition according to the GNW theory of consciousness[17,18]. Onset times were earlier than the predicted time of frontoparietal neural ignition

according to this theory[12], although they were still delayed relative to the appearance of category information in sensory regions, consistent with the idea that content selective activity in frontoparietal regions only emerges after activity in sensory areas reaches a critical level. The discriminating patterns were stable between ~200-600ms in both regions (Figure 4b, temporal generalization matrices; % of points with significant generalization ($q_{FDR} < 0.05$) between 200-600ms: PFC, 72%; Par, 93%). However, unlike posterior regions, category information in PFC and Par then subsided and was not sustained for the entire presentation of the stimulus (Figure 4c; AUC at 900ms (mean ± SEM across category comparisons): PFC, 55 ± 6%, Par, 57.2 ± 2.8%, compared to VT, 79.2 ± 9.6% and Occ, 77.2 ± 4.3%; PFC and Par all $p_{perm} > 0.05$ (uncorrected) except watch-object in parietal cortex ($p_{perm} < 0.012$ uncorrected)).

Previous studies have shown that HFA activity in the proximity of the orbit may be confounded with saccade related activity from extraocular muscles[50,51]. Importantly, these previous studies explicitly showed no contamination of responses in orbitofrontal cortex (as opposed to the temporal pole)[50], and patients were instructed to fixate during our task. Nevertheless, given the emphasis on prefrontal activation in some theories of consciousness[18,22], we performed several control analyses to ensure that the ability to decode category information from PFC was not driven by eye-muscle activity. First, we examined decoding performance separately in OFC and LPFC, the latter being less likely to be affected by eye-muscle activity. Category information was significantly decodable from both regions, though it was more prominent in OFC relative to LPFC (Figure 4d-e, see Supplementary Figure 6c for the dynamics of decoding in all comparisons; peak AUC (mean ± SEM across category comparisons): OFC 83.8 ± 2%, LPFC 77.3 ± 2.3%). Second, we repeated the decoding analysis in OFC using a bipolar montage, which emphasizes local neural activity as opposed to more distant ocular signals (Supplementary Note 1a). Category information remained significantly decodable from OFC (Figure 4f and Supplementary Figure 6d), and decoding time-courses were highly correlated between the average reference and bipolar montages (Figure 4g; mean ± SEM across category comparisons: $r(198) = 0.83 \pm 0.04$, range: 0.67–0.95). Finally, we examined decoding performance in OFC in 200ms time-windows without any saccadic events. Eye-tracking was not possible in the ICU setting, yet we were able to identify reliably the occurrence of saccadic spike potentials in one patient using an electrode placed over the temporal pole (posterior to the right eye; Supplementary Note 1b[9,52]). Despite the reduction in the number of exemplars (across categories and time-windows 11–43% of exemplars were excluded due to the presence of saccades, mean ± SEM: 24 ± 3%), decoding performance survived (Figure 4h black circles $p_{perm} < 0.05$ Bonferroni corrected for the number of windows), and was comparable to decoding results using all trials (mean difference ± SEM across categories and time-windows: 0.1 ± 1.3%; correlation: $r(22) = 0.8$, $p < 0.0001$; see Supplementary Figure 6e for more details).

To conclude, we found significant category information in PFC, which is not driven by oculomotor activity. Importantly, there was no overt report for any of the stimuli used in this analysis, excluding the possibility that this result stems from response-related signals[29,53].

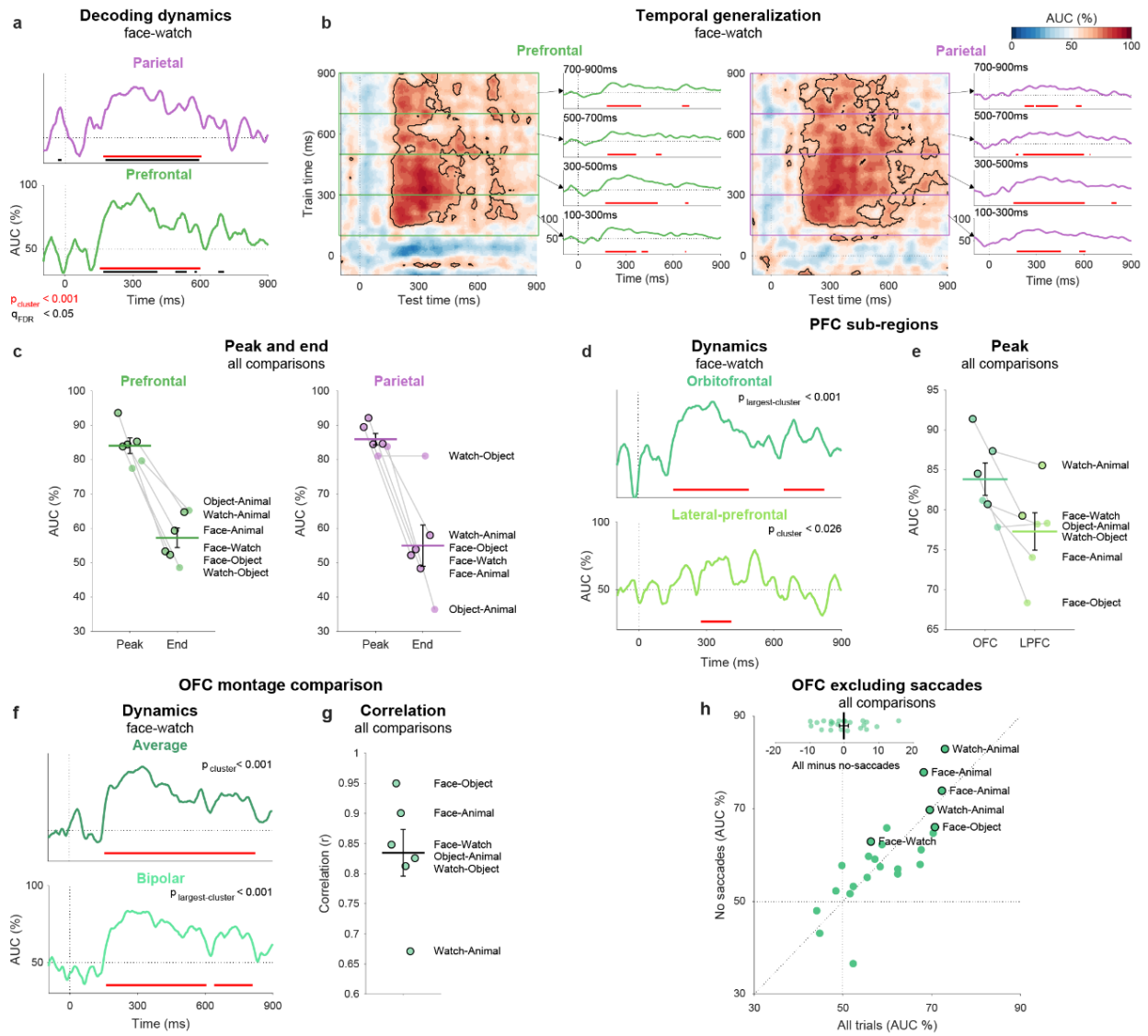

Figure 4. Category information is decodable from parietal and prefrontal cortex and is not the result of saccadic events. All analyses in this figure use images presented for 900ms or longer. **a** Decoding AUC time-course in PFC and Par (face-watch comparison, notations and statistical testing as in Figure 3c), reveal reliable coding of category information outside of posterior sensory regions (for other pairwise comparisons see Supplementary Figure 4). **b** Temporal generalization matrices in PFC and Par show high category information at the onset, which then subsides (notations and statistical testing as in Figure 3d). **c** Comparison of peak decoding and decoding at 900ms for all pairwise category classifications in PFC and Par. Results corresponding to the same classification are connected by a gray line. The mean across classifications is denoted by a horizontal line, error bars mark ± SEM. Dots are outlined in black for classifications significant using cluster-based permutations (Supplementary Figure 4c-d). **d-e** Decoding in sub-regions of PFC (orbitofrontal, OFC; lateral-prefrontal, LPFC). Category information was present in both sub-regions, though it was higher and persisted longer in OFC. **d** Decoding time-courses for the face-watch comparison (see Supplementary Figure 6c for the other comparisons). Red horizontal lines mark significant clusters. Dashed lines mark stimulus onset and chance level (AUC = 50%). **e** Peak decoding for all pairwise comparisons for OFC an LPFC (notations as in **c**). **f-g** Comparison of decoding in OFC with average reference vs bipolar reference (Supplementary Note 1a). **f** Decoding time-courses for the face-watch comparison (notations as in **d**; see Supplementary Figure 6d for the other comparisons). Only grids fully located over OFC were used in this analysis, which resulted in minor differences between the average reference decoding depicted here and OFC decoding in **d**. **g** High correlation of average and bipolar decoding time-courses for all pairwise comparisons. Each dot marks the correlation for a different comparison, the mean correlation across comparisons is marked by a horizontal line, error bars mark SEM. **h** Similar decoding accuracies in time

windows with and without saccades in OFC (single patient where reliable saccade detection was possible, Supplementary Note 1b). Each dot stands for decoding results in a different comparison and different time window (out of four non-overlapping windows between 100 and 900ms). The analysis was conducted once using all trials (x-axis) and once using only trials without any saccadic event (y-axis). Chance level (AUC = 50%) and equality (x=y) are marked by dashed lines. Black circles mark the comparisons which were significant in the analysis which did not include any saccadic events ($p_{perm} < 0.05$, Bonferroni corrected for the number of windows). The inset shows the difference between AUC using all trials and AUC using only trials with no saccades, which was not significantly different from zero. The mean across categories and time-windows is marked by a vertical black line and error bars denote SEM. For more details see Supplementary Figure 6e.

Decoding of visual category in sensory regions, and not frontoparietal regions, tracks the ongoing percept

The previous sections focused on responses to stimuli presented for 900ms or longer, to examine sustained responses. We now ask whether the sustained decoding profile found in visual sensory regions corresponds to the ongoing visual percept, or it is merely reflecting a prolonged onset response, irrespective of the actual duration of the percept. To answer this question, we trained linear classifiers to decode the presented category for each stimulus duration separately (300, 900 and 1500ms). We focus here on the face-watch comparison, but similar results were obtained for the other comparisons (Supplementary Figure 7d). In agreement with our previous findings (Figure 3c and Figure 4a), there was significant category information in all ROIs, for all stimulus durations (Figure 5a; all $p_{cluster} < 0.002$; for technical reasons the analysis was performed separately on patients S1-S3 and S4-10 (Supplementary Figure 7a), for each region we present only one of the groups; results from the other group were highly similar and are shown in Supplementary Figure 7b-e). Category information in VT followed the presence of the stimulus – for all three durations significant AUC clusters emerged shortly after stimulus onset, persisted throughout stimulus presentation, and subsided with a delay of approximately 230ms after the offset (or longer for the short 300ms stimuli; cluster extent: 300ms, 55-755ms; 900ms, 60-1130ms; 1500ms, 45-1735ms; corresponding to the results of point-by-point permutation testing with FDR correction, Supplementary Figure 7c). Occipital cortex showed a similar pattern, with shorter offset delays (cluster extent: 300ms, 50-630ms; 900ms, 110-1025ms; 1500ms, 90-1445ms). These differences were statistically significant in a direct comparison between decoding time-courses for the 1500 and 900ms stimuli, and the time-courses for 900 and 300ms stimuli (Figure 5b). Both contrasts showed significant differences in Occ and VT (cluster extent: $VT_{900-300}$, 640-900ms, $p_{cluster} < 0.002$; $VT_{1500-900}$, 1105-1500ms, $p_{cluster} < 0.001$; $Occ_{900-300}$, 535-900ms, $p_{cluster} < 0.001$; $Occ_{1500-900}$, 1020-1165ms and 1190-1445ms, $p_{largest-cluster} < 0.006$). In contrast, decoding in parietal and prefrontal cortex did not correspond to the duration of the stimulus (Par_{300} 150-605ms, Par_{900} 80-475ms, Par_{1500} 145-525ms; PFC_{300} 155-595ms, PFC_{900} 150-640ms, PFC_{1500} 155-440ms; difference between decoding time-courses all $p_{cluster} > 0.2$). This conclusion is also supported by decoding of other category comparisons (Supplementary Figure 7d) and by decoding using only electrodes which increased activity relative to baseline (positive responses; Supplementary Figure 7e).

Examining the temporal generalization matrices supported these conclusions (Figure 5c-d for VT and PFC; see Supplementary Figure 7f-g for Occ and Par) – all regions showed a stable pattern of decoding during the time-period when category information was found, but only in posterior regions it corresponded to the duration of the stimulus ($Occ_{900-300}$ $p_{cluster} < 0.016$, $Occ_{1500-900}$ $p_{cluster} < 0.002$; $VT_{900-300}$ $p_{cluster} < 0.008$, $VT_{1500-900}$

$p_{cluster}$ < 0.005; $Par_{900-300}$ and $Par_{1500-900}$ not significant; $PFC_{900-300}$, $p_{cluster}$ < 0.011, yet $PFC_{1500-900}$ was not significant, arguing against duration tracking in the region).

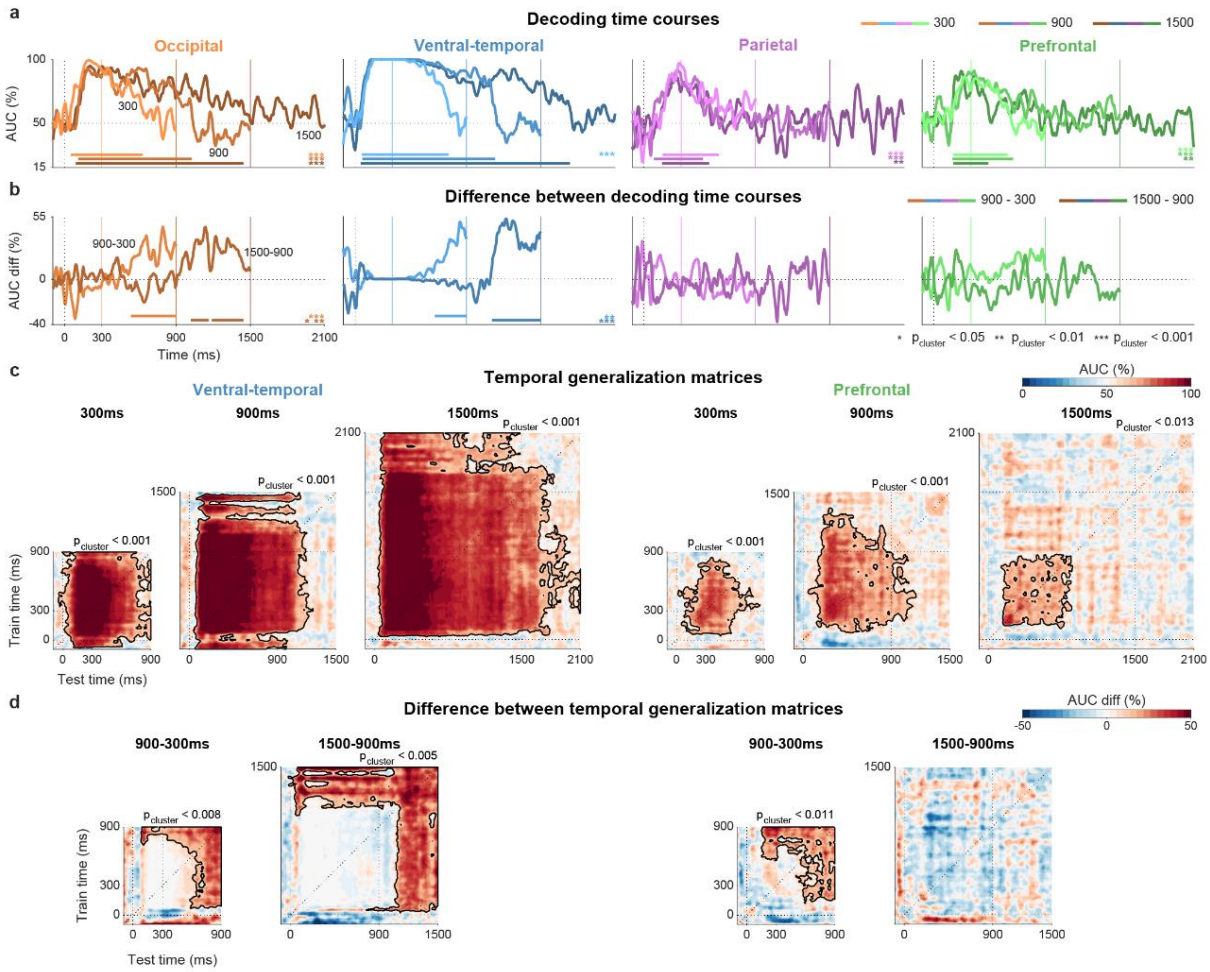

Figure 5. Category information in posterior regions, but not frontoparietal regions, tracks the duration of the stimulus. a Decoding dynamics in the four main ROIs separately for each stimulus duration (300, 900 and 1500ms; darker lines correspond to longer stimuli), for the face-watch comparison (see Supplementary Figure 7d for other comparisons). Occ, VT, and PFC are shown from patients S4-S10 (see Supplementary Figure 7b for Occ and VT in patients S1-S3; there were no PFC electrodes in S1-S3), Par is shown from patients S1-S3. Results are shown until 600ms after stimulus offset for each duration, since this was the shortest ISI and beyond that point another stimulus could be presented. Stimulus offset is marked by a vertical line of the same color. Chance level (AUC = 50%) and stimulus onset are marked by black dashed lines. Significant clusters of above chance decoding (by cluster permutation analysis) are marked by horizontal bars of the same color. Cluster p-values are indicated by the asterisks at the bottom right side of each plot (corresponding to the cluster temporal order). Similar time-ranges were found with point-by-point permutation and FDR correction (Supplementary Figure 7c). **b** Difference between decoding time-courses of 1500 and 900ms stimuli (dark line) and 900 and 300ms stimuli (bright line). Significant differences in decoding accuracy between durations (when one stimulus ends and the other continues) show that decoding in VT and Occ corresponds to the ongoing experience, but not in Par or PFC. Statistical testing and notations as in **a**. Only time-points after the offset of the shorter stimuli (where differences were expected) were considered in the calculation; results are nearly identical considering the entire presented range. **c** Temporal generalization matrices (see Figure 3 caption) for each stimulus duration in VT and PFC (see Supplementary Figure 7f-g for the other regions). Stimulus onset and offset, and the diagonal (corresponding to the lines in **a**) are marked by dashed lines. Statistical significance was assessed using cluster-based permutations; significant clusters are surrounded by black contour lines and the corresponding p-values are in

the top-right corner. **d** Comparison between durations (900 and 300ms, 1500 and 900ms). All time-points were considered in the calculation. For the number of exemplars viewed in each duration see Supplementary Table 6.

Information about specific images is maintained reliably across the cortex

Perception is more than recognizing the category of an image. For example, when we look at a face, we don't simply see a face, we see a specific instance of a specific face. We therefore turned to examine exemplar-level information, using representational similarity analysis (RSA)[54,55]. RSA captures the neuronal representational structure by examining the pairwise distances (or dissimilarities) between neural responses to each pair of stimuli, grouped together in a representational dissimilarity matrix (RDM). We reasoned that single exemplar information is maintained reliably if the representational structure, conveyed by the dissimilarities, is preserved across separate repetitions of the same exact stimuli. Thus, for this analysis we used only stimuli that were viewed at least twice by the same patients (focusing on images presented for 900ms or longer, as in previous analyses). We focus the analysis on five patients who viewed the same set of 60 images at least twice (194 responsive electrodes, Supplementary Figure 8a), as another four patients viewed less than 30 images twice, and one patient did not view any images twice. Results from a larger group of eight patients (with only 18 shared images) are shown in Supplementary Figures 8-9.

We defined complementary measures of exemplar-level information, Item reliability and Geometry reliability, both calculated on a time-point by time-point basis, for each region separately. Item reliability measures whether the position of each stimulus within the overall geometry of responses remains stable across repetitions (Figure 6a), and Geometry reliability captures global aspects of the representation by comparing the full dissimilarity structure across repetitions (Figure 6b). Example data from occipital cortex is shown in Figure 6c. Note that in both cases we compare dissimilarities within and between repetitions, which ensures not only preservation of the dissimilarity structure between repetitions, but also a similar state-space location. Additionally, Geometry reliability (and to a lesser degree Item reliability) is also sensitive to whether exemplars are distinguishable within the geometry, since we standardize the measures using a null distribution obtained by shuffling stimulus identity. Therefore, significant Geometry reliability indicates both representational reliability and discriminability of single exemplars within the representation. Conversely, if a geometry only represents distinctly some of the exemplars, we may fail to find significant Geometry reliability even when the geometry is fully maintained across repetitions. For a more detailed explanation the reader is referred to the Methods section.

We examined exemplar-level information using both correlation dissimilarity and Euclidean dissimilarity, as these are sensitive to different aspects of the response[56]. Both dissimilarity metrics showed widespread significant Item reliability (Figure 6d; for single patients see Supplementary Figure 8d). Across all ROIs peak Item reliability was more than three and a half standard deviations (SD) above the mean of the null distribution, and it was more than double this amount in posterior sensory regions (max-statistic permutations, $p_{perm} < 0.001$ in all cases, except PFC using correlation dissimilarity where $p_{perm} < 0.014$). Within PFC, both OFC and LPFC showed significant single-Item reliability using both Euclidean and correlation dissimilarities (all $p_{perm} < 0.003$; Figure 6e). Using only electrodes which increased their activity relative to baseline (Supplementary Figure 8b), or using data from the larger patient group with 18 images

(Supplementary Figure 8c) elicited similar results. We found significant Geometry reliability in occipital and temporal cortices (VT and Occ, $p_{perm} < 0.006$; LT, $p_{perm} < 0.016$; Figure 6f). Geometry reliability was not significant for PFC as a whole (although looking separately at PFC subparts, we found significant Geometry reliability in LPFC using Euclidean dissimilarity, $LPFC_{Euc}$ $p_{perm} < 0.014$, Figure 6g).

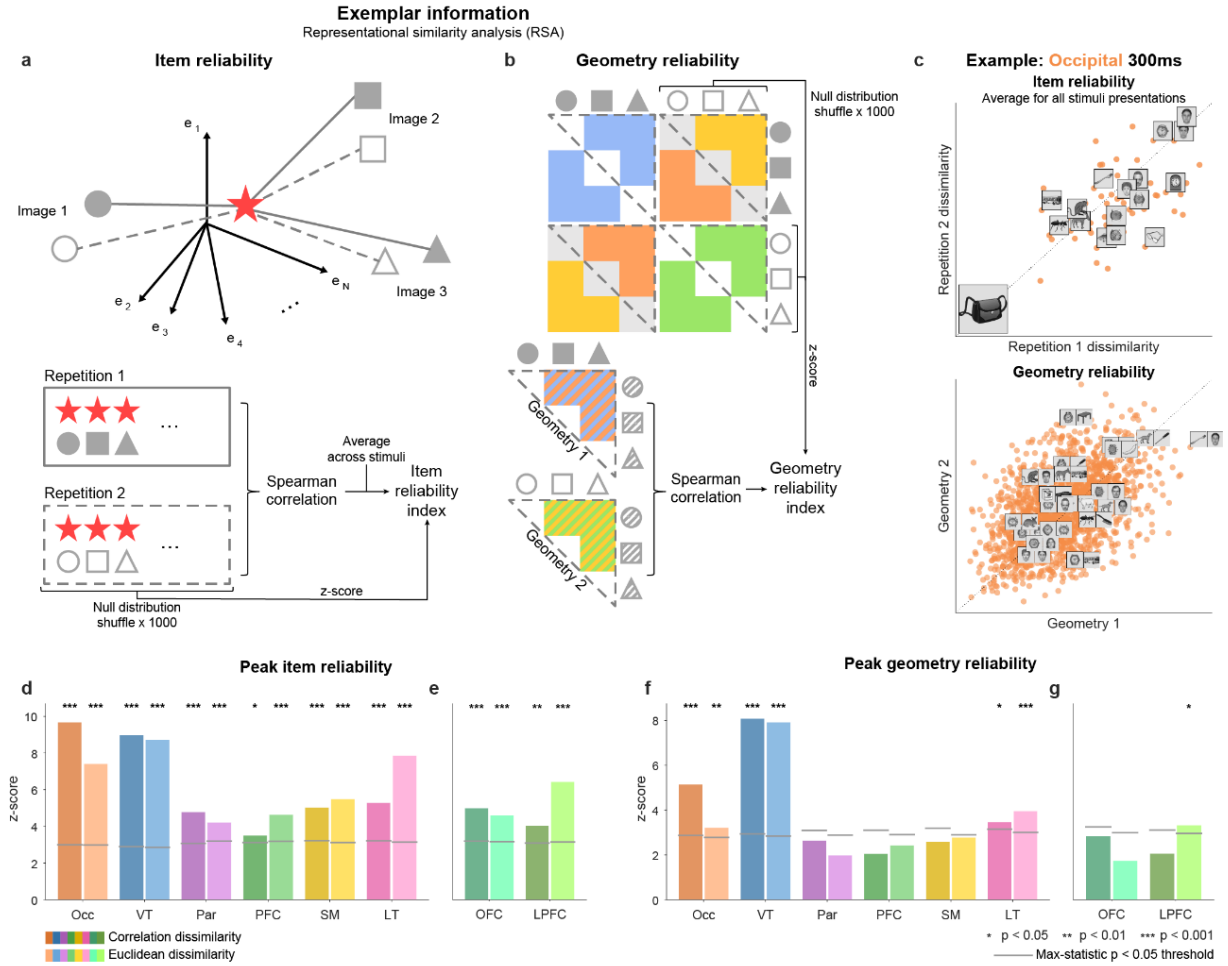

Figure 6. Specific images are represented reliably in widespread areas across the cortex. a Item reliability (schematic): For each stimulus presentation separately (example image presentation marked by a red star) we compute the distance (dissimilarity) in neural state-space of the current image to all *other* images from both repetitions (denoted by empty and filled markers; exemplar identity is denoted by the specific shape: circle, square, triangle). We then correlate the vector of dissimilarities from images in repetition 1 with the dissimilarities from the same images in repetition 2. This process is repeated for every image in both repetitions, and the results are averaged and z-scored with respect to the null (permutation) distribution (see Methods). **b** Geometry reliability (schematic): We first compute the dissimilarities between all stimulus presentations (all images, both repetitions). This results in a symmetric matrix composed of four distinct representational structures (top): dissimilarities within repetition 1 (blue), dissimilarities within repetition 2 (green), and two sets of dissimilarities between repetitions (yellow and red). Second (bottom), we average one within repetition structure and one between repetition structure (dashed triangles; forming Geometry 1 and Geometry 2), unfold them into two vectors and correlate. Finally, we z-score the resulting correlation (similarly to **a**). Averaging different dissimilarity structures or correlating each pair separately and then averaging the results led to similar results. Both measures were computed on a time-point by time-point basis. All analyses use images presented for 900ms or longer at least twice (faces, 18; watches, 19; objects, 13; animals, 10), data from a subgroup of five patients. **c** Example data for the computation of both measures in Occ, 300ms after onset. Each dot depicts one inter-exemplar dissimilarity (neural distance between two specific exemplars), compared across different

exemplar presentations. Dots closer to the equality (x=y) line indicate more reliable representation of exemplar-level information. Top (Item reliability): dots represent dissimilarities to one specific image response (in this example, one of the presentations of the handbag shown at the bottom-left corner), which are compared across repetitions. Bottom (Geometry reliability): dots represent all pairwise dissimilarities, which are compared simultaneously (images of some of the pairs are shown for illustration). **d-e** Peak Item reliability index, computed using correlation dissimilarity (dark bars) and Euclidean dissimilarity (bright bars), per ROI. We find wide-spread exemplar-level information across the cortex. Statistical significance by permutation test comparing to the peak Item reliability index of each permutation (max-statistic control for multiple comparisons). Gray horizontal lines denote the threshold for significance (95% of permutation values). **f-g** same as **d-e** but for geometry reliability.

Representation of exemplar identity is sustained and stable in posterior sensory regions and transient in frontoparietal regions

Next, we turned to examine the temporal dynamics of exemplar specific information. In the remainder of the text, we focus on results obtained using correlation dissimilarity, unless there was a substantial difference between the two measures. Figure 7a-b depicts the time-point by time-point Item reliability and Geometry reliability indices in the four central ROIs for images presented for 900ms or longer (see Supplementary Figure 8e for dynamics in SM and LT). As with the decoding analyses, we first examine posterior sensory regions, followed by inspection of frontal and parietal regions.

Both Item and Geometry reliability indicated that representation of single exemplar information in posterior regions emerges quickly after stimulus onset, with significant clusters starting at 75ms in Occ, and at approximately 100ms in VT (Figure 7a-b). In VT both exemplar reliability measures persisted throughout stimulus presentation (clusters extending until 900ms, $p_{cluster} < 0.001$; significant also using time-point by time-point permutation testing with FDR correction for multiple comparisons[48], $q_{FDR} < 0.05$ for all points included in the clusters, see also the diagonal in Figure 7c). In occipital cortex, Item reliability was significant until 900ms ($p_{cluster} < 0.001$), but Geometry reliability was split into two significant clusters with a brief gap (75-630ms, $p_{cluster} < 0.001$ and 645-845ms, $p_{cluster} < 0.012$). In VT both reliability measures, and especially Item reliability, remained high throughout stimulus presentation (mean Item reliability after 100ms: z = 7.3 SD, 82% of peak value; mean Geometry reliability after 100ms: z = 5.6 SD, 69% of peak value; Figure 7d-e), yet this was not the case in Occ. Despite the fact that sustained response amplitudes in occipital cortex were higher than those in VT both in absolute terms (Figure 1c), and relative to the peak response in each region (less attenuated, Figure 1f), around 550ms exemplar information in occipital cortex dropped by more than 50% (mean 550-900ms relative to mean 200-550ms). We found similar results repeating the analysis using only positively responding electrodes (Supplementary Figure 8h), and information dynamics captured by Euclidean dissimilarity revealed an even stronger distinction between the two regions (Supplementary Figure 8g). It should be noted that with this patient sub-group our coverage of occipital cortex is less extensive than the coverage of VT (Supplementary Figure 8a), but a similar picture emerged using the larger eight patient group with the smaller exemplar set (Supplementary Figure 8i). Thus, we conclude that representation of exemplar information is highly sustained in VT, but it is less robust in occipital cortex.

Consequently, we asked whether exemplar information is dynamic, or stable, similarly to the coding of category level information (Figure 3d). To test this, we extended our reliability measures by correlating representational structures between time-points (followed by z-scoring as in the reliability analyses),

analogous to the temporal generalization decoding method. To ensure that we are measuring stability in exemplar representation, as opposed to spontaneous activity fluctuations (shown to exhibit strong temporal dependencies[57–59]), each dissimilarity was compared (across time) to a dissimilarity between the same pair of images, but where at least one of the image presentations (repetitions) differed (see Methods). Echoing our decoding results (Figure 3d), Item-reliability information was remarkably stable in VT, maintaining very high correlations between representational structures in different time-points (Figure 7c). Mean Item stability (off-diagonal) between time-points after 100ms was more than 87% of the mean reliability index (on-diagonal values) during this period (Figure 7d). Item stability in occipital cortex showed a similar pattern (Figure 7c; off-diagonal to on-diagonal ratio (mean ≥ 100ms): 81.7%), though both mean on-diagonal and mean off-diagonal were attenuated relative to the onset reliability (Figure 7d). Examining Geometry stability led to highly similar results (off-diagonal to on-diagonal ratio (mean ≥ 100ms): VT, 84.7%; Occ, 84%; Figure 7e, see Supplementary Figure 9b for the full matrices). Repeating the analysis with Euclidean dissimilarity or with the larger patient group replicated these findings in VT (Supplementary Figure 9c-e). In Occ the results were less robust: we found stable representation (significant generalization) only using positively responding electrodes, but not with Euclidean dissimilarity or with the large patient group (Supplementary Figure 9c-e).

We found exemplar information also in other cortical regions, including parietal and prefrontal cortex. Item reliability was significantly above chance in both regions (Figure 7a), emerging first in parietal cortex (two significant clusters, 195-420ms, $p_{cluster} < 0.001$ and 445-540ms, $p_{cluster} < 0.013$) and then in PFC (385-465ms, $p_{cluster} < 0.015$). Repeating the analysis using Euclidean dissimilarity, using only positively responding electrodes or using electrodes from the larger patient group revealed a similar, and even stronger pattern (Supplementary Figure 8g-i; Par, all $p_{largest\text{-}cluster} < 0.007$; PFC, all $p_{largest\text{-}cluster} < 0.005$). Within PFC, we found significant Item reliability clusters in both OFC and LPFC (OFC, 195-260ms, $p_{cluster} < 0.033$ and 375-430ms, $p_{cluster} < 0.019$; LPFC, 385-445ms, $p_{cluster} < 0.04$; Supplementary Figure 8f). Geometry reliability was marginally significant in parietal cortex by cluster-based permutations (Figure 7b, 205-290ms, $p_{cluster} < 0.043$) but not with the stricter max-statistic approach (Figure 6f). Using positive responses alone we found significant Geometry reliability (Par $p_{cluster} < 0.002$; Supplementary Figure 8h; max-statistic peak $p_{perm} < 0.01$) but it was not significant using electrodes from the full patient group (Supplementary Figure 8i). Geometry reliability was not significant in prefrontal cortex. We conclude that single item information is represented outside of the visual cortex and the ventral visual stream, though it is not maintained as reliably as in posterior regions and subsides soon after the onset response (see Supplementary Figure 9a for stability between time-points). The dissociation found between Item reliability and Geometry reliability suggests that some exemplars are represented reliably, but the representation does not distinguish all exemplars, a prediction which we confirm in the next section.

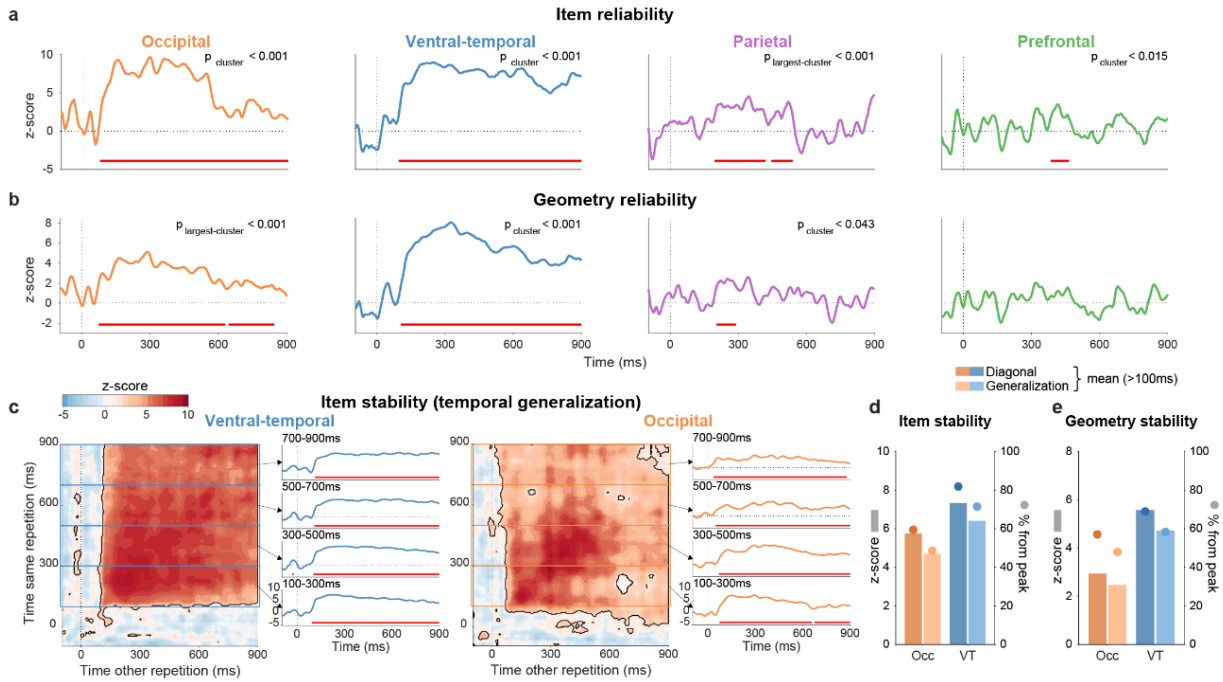

Figure 7. Exemplar specific information is sustained and stable in posterior sensory regions and transient in frontoparietal regions. a Item reliability and **b** Geometry reliability dynamics show that single item information is reliably maintained in VT and Occ throughout the presentation of the stimulus, and only transiently maintained in Par and PFC (see Supplementary Figure 8e for the remaining regions). Dashed lines mark stimulus onset and chance level (no single-item information). Red horizontal bars mark significant clusters by cluster permutations (the corresponding cluster p-value is shown in the top-right corner). **c** Item representation stability between time-points in VT and Occ. The diagonal of these matrices corresponds to the dynamics time-courses in **a**. Representation stability between t_1 and t_2 was computed by correlating for each stimulus presentation the vector of neural dissimilarities to the rest of the images in the *same* repetition at time t_1 (y-axis) with the dissimilarities from the *other* repetition of these images at time t_2 (x-axis). Contiguous points with $q_{FDR} < 0.05$, by permutation test and FDR correction for multiple comparisons are surrounded by black contour lines. Insets on the right depict the mean Item stability index between 200ms windows (100-300ms, 300-500ms, 500-700ms or 700-900ms) to all time-points, corresponding to the outlined regions of the matrix. In the insets on the right red horizontal bars mark points which were significant for at least 50% of time-point in the averaged range. **d** Comparison of Item reliability within the same time-point (matrix diagonal, darker bars and dots) to generalization (stability) of representation between time-points (matrix off-diagonal, lighter bars and dots). Bars denote the mean value (100-900ms) relative to the null distribution (z-score units; left y-axis), and dots denote the relative value (% from peak; right y-axis). In both regions stability between time-points was only slightly reduced from reliability of representation within the same time-points. **e** Same as **d** but for Geometry reliability (see Supplementary Figure 9b for the full generalization matrices).

Exemplar-level information persists after accounting for category structure

As we found robust coding of category information in both posterior and frontoparietal regions (Figure 3-5), it is theoretically possible that the reliability and stability of the representational geometry described in the previous sections reflects the categorical structure rather than item level structure. We first asked whether the representational dissimilarity structure captured by RSA indeed contains visual category representation. To this end, we designed four models of category information (Figure 8a). All models assume exemplars within a category are similar to each other, and dissimilar to exemplars in other categories. In some models we add a hierarchy of similarities between categories (supported by previous studies[60–62]), such that categories closer on the hierarchy are assumed to be more similar to each other. The models are: (1) Single-

category model – no relation between categories, only a primary category distinction. (2) Low-level model – faces and watches form one high-order category (both are photos of round objects), and other (non-watch) objects and animals form the second (pointy illustrations). (3) Semantic model (based on animacy) – faces and animals form one high-order category (animate) and all objects (watches and non-watches) form the second (inanimate). (4) "Face-vs-rest" – faces are distinct from all other categories, which are all grouped into one high-order category. We compared each of these models to the neural representational geometry (mean of the four RDMs in Figure 6b) using Spearman correlation[63]. The peak correlation with each of the models is shown in Figure 8b (Supplementary Figure 10a shows the full dynamics of the RDM-model correlations; results were highly consistent across time-points). Unsurprisingly, the representational geometry in occipital cortex was most correlated with the "Low-level" model ($\rho_{peak} = 0.47$) and in VT with the "Face-vs-rest" model ($\rho_{peak} = 0.73$), closely followed by the "Semantic" model ($\rho_{peak} = 0.72$). For parietal cortex we found the strongest correlation to the "Face-vs-rest" model ($\rho_{peak} = 0.27$), in agreement with the state-space trajectories (Figure 2a). The prefrontal RDM was only weakly correlated with any of the models. The strongest correlation was to the "Low-level" model, but it was modest ($\rho_{peak} = 0.09$) and did not survive correction for multiple comparisons (max-statistic approach). We attribute the difference between this finding and our decoding results (Figure 3b) to the partial coverage, stemming from the exclusion of patients which did not view enough exemplars twice (Supplementary Figure 8a). Thus, single-item information found in PFC is unlikely to be attributed to category affiliation, but for completeness we include PFC in the following analyses as well.

The high correlations observed in VT and Occ (and to a lesser extent Par) reveal the neural RDM contains category information, in agreement with previous studies[60–62] and with our own previous results (Figures 2-4). However, it also shows that a substantial proportion of the information within the RDM is *not* captured by categorical structure (mean unexplained variance ($1-\rho^2$) after 100ms: Occ, 91%; VT, 78%; Par, 98%; PFC, 99.8%). We next asked whether this unexplained information is reliable and stable. To this end, we repeated the calculation of both reliability measures using partial correlation to remove the category structure conveyed by the model RDM. Figure 8c depicts the resulting Item reliability dynamics after partialling-out the category model which was *most* strongly correlated with the neural RDM, thereby providing the strictest measure of category information (using other models led to similar or higher reliability scores, Supplementary Figure 10b). For comparison, the gray lines depict the Item reliability scores without removing category information (Figure 7a). Despite some reduction in the fidelity of representation, Item reliability remained significant in all regions after partialling-out category information (Occ and VT $p_{cluster} < 0.001$, Par and PFC $p_{cluster} < 0.024$). Moreover, it remained sustained and stable in VT, though not in Occ. In VT we observed significant partialled Item-reliability in a cluster from 95-900ms, also supported by point-by-point permutation testing with FDR correction (100% of points after 100ms showed significant reliability, $q_{FDR} < 0.05$; Figure 8d). We found significant generalization (off-diagonal, stability between time-points) between %80 of time-points after 100ms (93% focusing on points after 200ms). In comparison, after partialling-out category information from Occ Item reliability was only significant in a cluster from 80-615ms (with similar conclusions from point-by-point tests, Figure 8d; see Supplementary Figure 10c for the

full matrices from both regions). Repeating this analysis with Geometry reliability led to highly similar results (Supplementary Figure 10d-e). Thus, we find reliable exemplar representation in anterior and posterior regions, and sustained and stable representation in VT, which is not accounted by the category information expressed by our models (Figure 8a), indicating that exemplar-level information is maintained over and above the category distinctions.

As a final test, we examined exemplar level information within specific categories, which is not be influenced by the category structure. To account for the substantial reduction in the number of exemplars (between 3-fold and 6-fold reduction, depending on the category), we smoothed data with a 100ms moving window prior to this analysis (see Supplementary Figure 11a for results without this step). Item reliability was significant in all ROIs for at least one category (peak Item reliability in the four main ROIs is shown in Figure 8e; see Supplementary Figure 11b for SM and LT), indicating single exemplar information is coded in these regions independently of category distinctions. This can also be gleaned from Figure 8f, which depicts the neural responses to all image presentations from one example category at the time of peak reliability, revealing more similar neural responses to repetitions of the same image compared to responses to different images. Occ and VT showed significant Item reliability for faces, objects and animals (all $p_{perm} < 0.008$, peak reliability $z > 3.5$ SD). Unsurprisingly, no region showed reliable representation of single watches, as this category was designed to be highly homogeneous (Figure 1b). Par showed marginally significant Item reliability for faces and objects ($p_{faces} < 0.013$ and $p_{objects} < 0.021$); however, unlike the other findings, this result did not survive Bonferroni correction for four categories and should therefore be taken with caution. In PFC we found reliable representation of face exemplars ($p_{perm} < 0.002$, peak reliability $z = 4$ SD) and not of any of the other categories (although animals were significantly represented in OFC, Supplementary Figure 11c). Thus, we conclude that VT and Occ represent exemplar information across multiple semantic domains, irrespective of category information, and PFC represents information about specific faces. These conclusions are also supported by examination of the full temporal dynamics (example time courses shown in Figure 8g, see Supplementary Figure 11d-e for other categories and stability in time matrices).

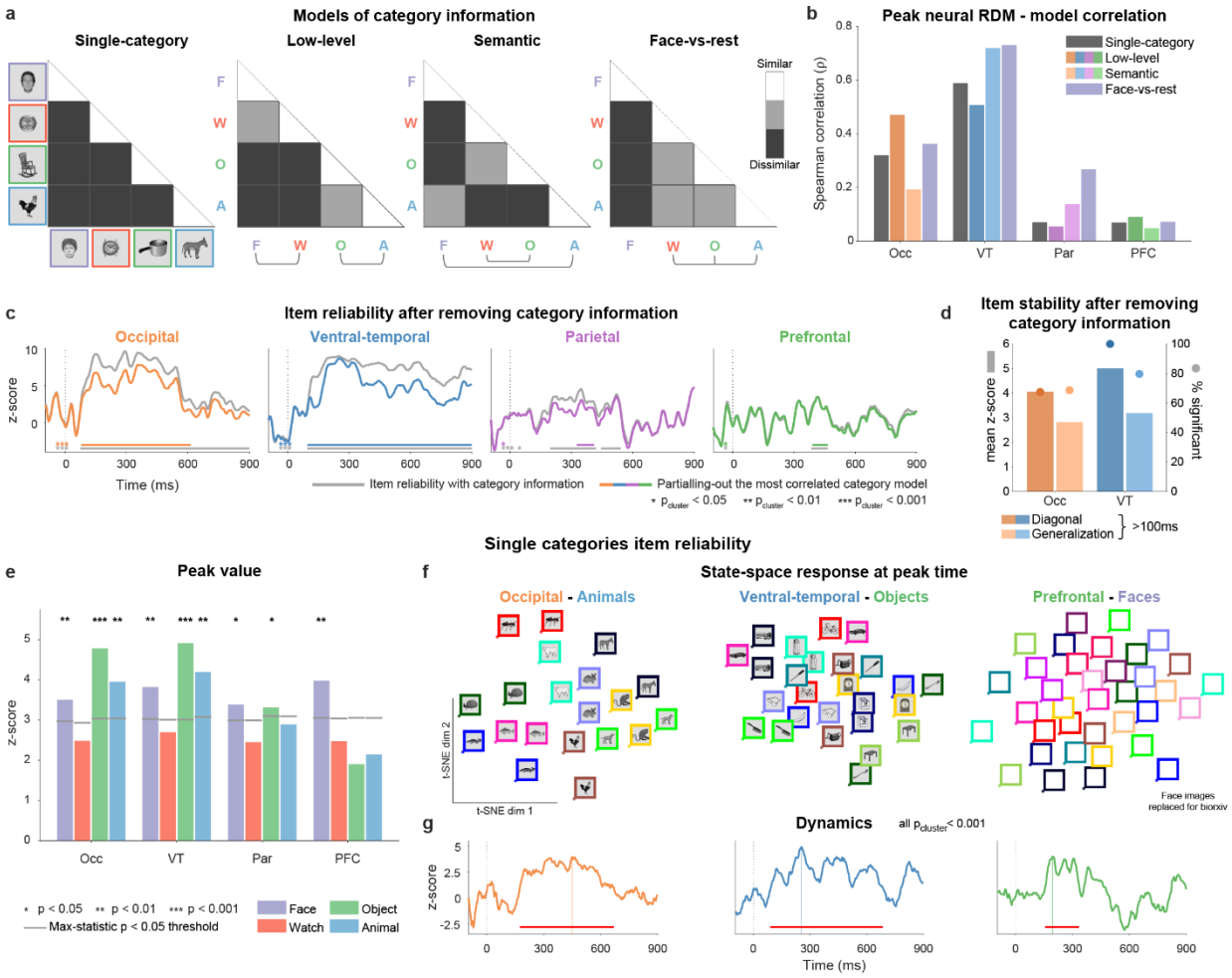

Figure 8. Identity of single images is represented reliably beyond the visual category of the image. a Models of potential category information captured by the representational dissimilarity matrix (RDM). In all models, exemplars within each category are modelled as 100% similar to each other, and dissimilar from exemplars in all other categories (0% similarity). In three of the models, we added a hierarchical category structure, such that categories closer on the hierarchy are assumed to be more similar to each other (50% similarity). From left to right: (1) "Single-category" model – no relation between categories, only a primary category distinction. (2) "Low-level" model – faces and watches form one high-order category (both are photos of round objects), and other (non-watch) objects and animals form the second (pointy illustrations). (3) "Semantic" model (based on animacy) – faces and animals form one high-order category (animate) and all objects (watches and non-watches) form the second (inanimate). (4) "Face-vs-rest" – faces are distinct from all other categories, which are all grouped into one high-order category. Only the lower triangle is depicted as the matrices are symmetric. **b** Peak correlation between the neural RDM and the model RDMs, for the four main ROIs. VT and Par were most correlated with the "Face-vs-rest" model, Occ and PFC were most correlated with the "Low-level" model. See Supplementary Figure 10a for the full temporal dynamics (including SM and LT). **c** Item reliability dynamics after partialling-out category information. Colored lines depict reliability scores after partialling out the category model which was *most* correlated with the neural RDM (see Supplementary Figure 10b for the other models). For comparison, gray lines depict the original Item reliability values, computed without partialling-out information (identical to Figure 7a). Significant clusters (cluster permutation analysis) of above chance reliability are marked by horizontal bars of the same color. Cluster p-values are indicated by the asterisks at the bottom left side of each plot (corresponding to the cluster temporal order). Dashed lines mark stimulus onset and chance level (no single-item information). Item-reliability was attenuated by the removal of category structure but remained significant in all regions. **d** Item stability after removing category information (using the strictest category model, as in **c**) – on-diagonal reliability (within the same time-point, dark shade) and off-diagonal stability (between time-points, bright shade) in VT and Occ (both after 100ms). Bars (left y-

axis) denote the mean value relative to the null distribution (z-score units), and dots (right y-axis) denote the percentage of significant points using point-by-point permutation testing with FDR correction on the full stability matrix (note this is different than the notation in Figure 7d-e). The full stability matrices for all models of category information are shown in Supplementary Figure 10c. **e-g** Single item information is present within in single categories. **e** Peak Item reliability in single categories (for SM, LT and a subdivision of PFC see Supplementary Figure 11b-c). Statistical significance assessed using max-statistic permutations to control for multiple comparisons. Gray horizontal lines denote the threshold for significance (95% of permutation values). All regions showed significant Item reliability in at least one category (though Par_{faces} and $Par_{objects}$ do not survive Bonferroni correction for four categories). **f** Visualization of neural population responses to two presentations of all images from one example category, during the time of peak reliability (Occ, animals at 450ms; VT, objects at 255ms; PFC, faces at 195ms) shows repetitions of the same exemplar are represented more closely than repetitions of different exemplars. The location of each image corresponds to the population response to that image. Image frames are color-coded according to exemplar identity to ease comparison between repetitions (not presented during the task; color code generated using 'distinguishable_colors'[64]). Population responses were projected to two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE)[65] with correlation dissimilarity (first 50 principle components, or the number of electrodes if lower). **g** Item reliability dynamics for the regions and categories presented in **f**. Dashed lines mark stimulus onset and chance level (no single-item information). Colored vertical lines on the abscissa indicate peak times. Statistical significance was assessed using cluster permutations (red horizontal bars mark significant clusters).

Discussion

Understanding the neural correlates of conscious experience is one of the most coveted yet most challenging goals of cognitive neuroscience (and some would say science at large)[10,11], leading to multiple competing theories[12,13]. In different guises, quest for the NCC typically involves looking for an isomorphism between a specific experience and neural signal by contrasting two states: having an experience (being consciously aware) of a stimulus and not having one[53,66]. Usually, this requires drastic and unnatural manipulations the of the stimuli, for example by different forms of masking, by or strong manipulations of attention[67]. Under these liminal conditions, observers sometimes experience stimuli and sometimes not, and neural signals can be compared. These are powerful paradigms, especially for examining processing without conscious awareness, but in many cases determining whether a stimulus was genuinely not experienced is difficult and when experience is present, it is typically impoverished due to the manipulation[68]. We suggested that temporal manipulations of the duration of visible stimuli provides an illuminating alternative, as neural correlates of experience should be present not only at the onset of experience (which previous studies examined) but be present as long as the experience persists ([9], see also refs [12,69]). By examining which aspect of the neural response follows the duration of experience, we addressed directly a glaring gap in the quest for the NCC, and more generally in vision neuroscience – how do we maintain a stable experience beyond the immediate onset response, when neural activity, as well as single site selectivity drops dramatically[8,9]?

Using intracranial data from ten patients, we find substantial variability in single electrode responses (Figure 1) and state-space trajectories (Figure 2) over time, including a five-fold reduction in the level of activation. Yet, despite this variability, we show that the pattern of activation across recording sites in occipitotemporal cortex contains sustained information about the visual percept at the category (Figure 3) and exemplar level (Figure 6-8), which is both highly stable (invariant) in time and lasts for the entire the duration of the stimulus (Figure 5). These properties correspond directly to our ongoing perceptual experience, positioning this information signal as a viable candidate for a neural correlate of conscious visual awareness. Additionally, despite the fact that no response was needed for any of the analyzed trials, we found an "ignition" of visual

information in frontoparietal cortex, including the PFC (Figure 4, Figure 6), which lasted for several hundred milliseconds after onset, but did not correspond to the full duration of the percept. This suggests that frontal regions are involved in updating perceptual experience. As we discuss below, these results shed important light on the claims of central theories of the neural correlates of conscious awareness.

Our results imply that it is not the level of activation at any given site that allows the stable experience, but the pattern of activation across multiple sites, within occipital and mainly ventral-temporal regions. Put differently, we show that while the population response as a whole is highly variable, part of it is invariant in time. The notion that perceptual representations in humans depend on multivariate representations rather than on local activity has been supported in the past mainly based on by fMRI findings[60,70]. By examining the stability of representation over time, our study extends these previous claims and provides strong support to the importance of multivariate patterns, echoing a broader transition in neuroscience to the "population doctrine"[41–43]. Both our category-level and our exemplar-level results reveal a subspace within the vast space of possible neural responses, which maintains the distinctions between different visual categories and different image identities in a time-invariant manner[55,71]. Thus, we hypothesize the existence of an "Experience Subspace", and it is the projection of the population response to this subspace which corresponds to our perceptual experience. The stability within the "Experience Subspace" also affords downstream regions with a stable readout of information about the visual percept[72,73].

Another way to conceptualize the "Experience Subspace" is through the lens of representational similarity analysis (RSA), which has emerged in recent years as a powerful approach to understand neural coding[54,55,63]. RSA quantifies representation not by looking at the responses to different stimuli in isolation, but by examining the (dis)similarity between responses to different stimuli, which together define a geometry of inter-stimulus relations. Recently, it was proposed that this relational geometry is constitutive of our phenomenal experience, that is, the properties and content of each experience are derived from the relation of its neural representation to other previously learned representations[19,74,75]. We confirm the representational geometry is a viable candidate for maintaining conscious perception by showing that the state-space position of each exemplar relative to other exemplars is reliable across repetitions (Item reliability) and that the geometry as a whole is maintained across repetitions (Geometry reliability). This was seen clearly in posterior regions, where the relational geometry was also stable in time, along the duration of a stimulus presentation, and it was also seen in PFC though to a lesser degree. Since each exemplar belongs to a category, the reliability of the neural geometry could merely reflect the categorical structure. However, our analysis showed that exemplar representational structure goes beyond the category representation, as it was stable after partialling out models of category information and persisted even within single categories. Thus, the stable experience of a particular exemplar over time may be maintained by the position of the response pattern it generates within the space of potential responses[19,74,75]. Because our stimuli set did not include the same objects in different contexts or from different viewpoints, these conclusions about are limited to viewpoint-specific images rather than to the independent identity of the stimulus (like the individual behind a face).

The independence of the pattern representation from the overall level of activation over time, which we demonstrated here, provides strong support for the above formulation. For the brain, at any given point in time only one of the categories is present, and online identification must be made based on the (nearly) instantaneous neural response to that object (with no comparison to what another stimulus would have elicited). Consider a site in the FFA with a strong onset response to faces and a smaller onset response to objects (say half), with little variability. There is an amplitude threshold which may yield perfect discrimination during the onset response, but the same threshold would yield poor discrimination later on, when both the response to faces and objects has decreased below the onset response for the weaker category. Vice versa, setting a threshold which would discriminate faces and objects 800ms after stimulus onset would perform poorly at the onset. Thus, setting a fixed threshold on a single site would not do. The brain could learn a different threshold of activation for every time point after stimulus onset. However, as our results show, using a fixed, learned pattern of relative activation across sites, provides the required information at any point after stimulus onset.

A perennial question in the philosophy of mind is whether a specific conscious experience has a one-to-one relationship to a specific physical (neural) state, or if the same conscious experience can be realized by more than one state (multiple realizability; for a more nuanced discussion see [76]). In theory, if multiple realizability is plausible, we could have found that the neural signals contain robust information about the experienced image at any given time along the duration of the stimulus, but the representing pattern at each time point is different. However, this was not the case. Instead, the same classifier could discriminate categories from the onset response to the end of the presentation time in occipitotemporal cortex, and similarly, the same representational geometry persisted throughout this time. This suggests that not only do we have an informative active representation of our visual percept as long as we experience it, but that this representation is to a large extent invariant. Thus, at least in this case, there is no need to invoke multiple realizability (see also [69]).

Our results have implications on central theories of NCC, which have long debated whether consciousness relies more on posterior, sensory regions of the brain or on the amodal frontal region[23,27]. The posterior representations found here seem isomorphic to the subjective experience of visual persistence, while the temporally limited frontal representations are more compatible with change-related responses. One could argue that observers are not aware of the stimulus throughout its duration, but only at its onset (and perhaps offset), and therefore the transient frontoparietal response is more akin to the conscious experience. By this account, the subjective experience of persistent awareness is illusory, a type of 'filling-in' phenomenon. While we cannot provide objective proof for the subjective experience, considering relatively short duration of the stimuli (up to 1.5 sec), the high visibility, and the fact that the task required the subjects to attend to the stimulus along its presentation, this seems unlikely to us, or at least, we think the onus in on those making the argument for filling in long durations. Another possibility is that our perception is actually composed of discrete steps, with less than half a second interval between samples[77,78]. If ignitions indeed take place every few hundreds of milliseconds, we may have missed them if they are not aligned between trials. However, this is true if the duration of an ignition is shorter than the interval between discrete samples, which does not seem to be the case (Figure 4, ref. [16]).

Alternatively, rather than being adversarial, the distinction between the persistent representation in occipitotemporal cortex and transient representation in frontoparietal cortex may map to the conceptual distinction between phenomenal and access consciousness[79,80]. The former pertains to the mere experience, not requiring any decision or report, whereas the latter refers to the more reflective part of acknowledging to oneself or others or making a decision about the presence of an experience. Phenomenal awareness requires a continuous instantiation, while access by nature is transient.

The recent COGITATE consortium has adopted (and adapted) our multi-duration paradigm and logic for one of the two key experiments in an adversarial collaboration aimed to adjudicate between two prominent theories of consciousness, the IIT and the GNW[12]. While that project is ongoing, a highly important contribution has already been made by delineating the predictions of the two theories regarding the NCC as it is reflected in the multi-duration paradigm[30]. Naturally, our analysis does not address all predictions, yet, as it stands, the results are quite compatible with predictions of both camps about the anatomical basis and temporal profile of activation, questioning the adversary nature of these predictions. Namely, IIT predicted that neural activity in the "posterior hot zone" including occipital and VT regions will persist, with a fixed representation, as long as the image persists (and presumably conscious experience persists) while it is basically agnostic about prefrontal activation. GNW predicted an initial onset response (ignition) in PFC, starting around 250-300ms after stimulus onset, reaching a peak and thereafter declining, without persisting representation. These predictions are in fact compatible rather than adversarial as they are not mutually exclusive, and indeed both predictions bore out in our study – posterior areas showed stable and persistent representation, tracking the duration of the stimulus, and PFC showed an ignition without persistence.

The findings that frontoparietal activations constitute a NCC has recently been criticized, suggesting that these activations do not reflect an experience per se, but the need to report it[29,53]. Indeed, studies that use no-report paradigms, where awareness is inferred indirectly without overt responses, found reduced participation of frontoparietal cortices[28,81]. The trials we analyzed here did not require any overt response from the subjects. Thus, the presence of frontoparietal activity cannot be related to action or reporting. However, unlike pure no-report paradigms, here the subjects had to decide whether the stimulus belonged to a target category or not, implying that our findings may still be related to task requirements. Nevertheless, our findings are unique in showing stimulus specific representations which are not mapped to task distinctions (see recent findings in monkeys in [82,83]). That is, we find highly accurate decoding between image categories that are both non-targets, and which do not share any low-level properties with the target stimuli (e.g., face and watches). Moreover, the task only required discrimination of category information, yet we find reliable exemplar representation, within trials that do not require any response. While this implies genuine perceptual information, not directly related to task demands or response selection, future studies should manipulate task demands directly.

Within PFC, we made a distinction between orbitofrontal (OFC) electrodes and lateral prefrontal (LPFC) electrodes as these parts of the cortex belong to somewhat different networks considering cytoarchitectonics, connectivity patterns and function[84,85]. While both regions evinced above chance decoding during the onset response, we found more robust evidence for visual category representations in OFC than in LPFC. Previous

studies suggested the "Frame-and-fill" model, stating that a 'quick and dirty' pathway transmits low spatial frequency information to the OFC, which in turn provides predictions, based on low resolution analysis, to ventral temporal regions, facilitating recognition[86,87]. However, considering that OFC category discrimination did not precede VT, our findings do not fully align with the prediction of this hypothesis.

The OFC is close to the orbits, raising the concern that signals in this region may be conflated with EMG signals from extra-ocular muscles (saccadic spike potentials[50,51,88]). There are several lines of evidence against the possibility of contamination of our OFC results by myogenic spike potentials. First, previous studies mapping the influence of this myogenic source on intracranial signals found significant spike potentials in temporo-polar electrodes, but not in OFC[50]. Second, we re-referenced HFA in OFC electrodes to neighboring electrodes, thus suppressing common, remote sources of signals, like the eye muscles, but above chance decoding endured. Third, in one patient we could use saccadic spike potentials in temporal pole electrodes to detect the occurrence of saccades and found above chance decoding in saccade-free segments. Taken together, it seems that the OFC decoding was not due to myogenic activity. Moreover, we found significant decoding in LPFC as well, and contamination by the spike potential is much less likely in this region.

The parietal cortex has been traditionally linked with the prefrontal cortex by theories of conscious awareness, based on findings that changes in perceptual content, like in binocular rivalry[89], or detection of liminal, masked stimuli (relative to no detection[25]), activate both parietal and frontal regions. Additionally, the parietal cortex, especially at the temporoparietal junction plays a role in other theories of conscious awareness as the attention schema theory[90]. Recently, the proponents of IIT suggested the concept of the "posterior hot zone". The exact anatomical boundaries are not precisely stipulated but crude diagrams and verbal descriptions of the hot zone typically include the parietal cortex with occipital and temporal cortex[24,91,92]. In the current results, the parietal cortex shows early representations of categories and exemplars, which do not persist for the duration of the stimulus. Thus, parietal responses resemble the PFC responses much more than the occipital and ventrotemporal cortex responses, in line with the prediction of the GNW. That said, the parietal lobe is far from being a homogenous structure[93], and quite likely different parts of the parietal lobe will have different response functions. This remains for further investigation.

Some limitations should be noted. An inherent weakness of ECoG studies is the sparse spatial sampling, determined solely on clinical indications. In our patient group, we had extensive coverage of occipitotemporal cortex and less dense coverage of frontoparietal regions. Thus, null results should be taken with caution as it is possible that other locations would have resulted in different conclusions. Following an extensive body of evidence showing that broadband, non-oscillatory high frequency ECoG activity is highly informative and is correlated with multiunit firing rate[35–37], we concentrated our analysis on this band. However, it is possible that other bands or even changes in synaptic efficacies underlie for example persistent representation in PFC[94,95].

Our parcellation of the brain to the 6 ROIs is admittedly only one of many possible choices and was chosen to conform to existing views about functional segregation in the brain. There is no claim here that for the brain, the useful activation or representation is constrained by the borders between these ROIs. We chose

this parcellation to adhere to common divisions of the brain, and to be able to address arguments from the different theories pertaining to these ROIs.

Future studies will need to examine whether there is any limitation for how long stable multivariate representations can persist. It is possible that under unchanging conditions, longer durations than the ones we tested will suffer more from internal or external distraction and examining the effects of such distraction will be an important venue for exploration. This may reduce however the ecological validity, as people do not tend to gaze for very long time at an unchanging object.

To conclude, by presenting diverse visual stimuli in different durations, we show that in occipital and ventral temporal cortex the relative pattern of activation in response to each stimulus is maintained reliably in relation to other stimuli, as long as the object is present, both at the category level and at the exemplar level, despite substantial attenuation in response magnitude during this time. Put differently, the population response as a whole displays rich temporal dynamics, yet, embedded in this variability we find stability within a novel type of invariant representational space, which is consistent with the subjective experience of a stable conscious percept. It is the invariant position within this "Experience Subspace" which imbues each perceptual experience with its unique quality. This may also explain the perplexing dissociation between perception of repeating stimuli as identical despite the ubiquitous phenomenon of repetition suppression[96,97]. In frontoparietal regions we find transient content representation, despite the lack of overt response. These findings are consistent with the recent predictions of both major theories of consciousness, IIT and GNW.

Methods

Participants

Ten patients undergoing presurgical evaluation for treatment of intractable epilepsy participated in the study (4 female, age (mean ± SEM): 41 ± 3.7, range: 19-65; for individual demographic details see Supplementary Table 1). Recordings were conducted in the epilepsy ICU. Seven patients were recorded in the Stanford School of Medicine, two in the California Pacific Medical Center (CPMC) and one in the University of California, San Francisco Medical Center. All patients gave informed consent approved by the University of California, Berkeley Committee on Human Research and corresponding IRBs at the clinical recording sites, in accordance with the Declaration of Helsinki. Results from the same dataset were previously reported in [9].

Stimuli and task

Patients viewed grayscale images, presented at the center of a uniform gray background, and extending approximately 5° of the visual field in each direction. Stimuli were presented on a laptop screen and responses captured on the laptop keyboard (Figure 1b, top). The images belonged to multiple semantic categories (Figure 1b, bottom), including faces (approx. 30%), man-made objects (watches: approx. 30%, other man-made objects: approx. 18%) and animals (approx. 10%). Approximately 10% of images were targets (see below). The remaining images (<3%) were mostly houses or body-parts, which were not analyzed due to paucity of exemplars (see Supplementary Table 4 for the number of stimuli viewed by each patient from each category). Watches and other man-made objects were considered separately in this study since

watch images, like the face images, were taken from a dataset of real photos, while the other objects were illustrations. When comparing to face images, watches are considered a better control over low-level similarities, as the two categories share the overall round outline with internal details[33,34], and for this reason watches were presented with the same probability as faces. To verify that information content tracked ongoing experience, images were presented for variable durations (Figure 1c; patients S1-S3: 300, 600, 900, 1200 or 1500ms, patients S4-S10: 300, 900 or 1500ms). The probability of each exemplar appearing in each duration was uniformly distributed. A fixation cross was displayed between image presentations (inter-stimulus interval: 600, 750, 900, 1050 or 1200ms). Note that the image sequence was randomized for each patient, meaning that specific exemplars were viewed for different durations by different patients.

Patients were instructed to fixate at the center of the screen and respond with a button press to the appearance of rare targets (Figure 1c). In the main experimental condition, performed by all patients, the target was any image of a clothing item. The second experimental condition was performed only by seven patients (S4-S10) in half of the blocks. In this condition patients were instructed to respond to two target types: (1) appearance of a clothing item (as in the main condition), (2) blurring of any image during the last 200ms of presentation (see [9] for more details on the motivation of the dual task). Both target types together comprised approx. 10% of the trials and were not analyzed in this study, focusing on no-report trials. Neural responses to non-targets did not differ between the conditions[9], therefore data from both conditions was analyzed together.

Data acquisition and preprocessing

Patients were implanted with 64-128 subdural electrodes (1004 in total), arranged in 1-dimensional strips and\or 2-dimensional grids (AdTech Medical Instrument Corporation). Electrodes were 2.3 mm in diameter, with 5 or 10 mm spacing between electrodes. Eight patients were implanted in the right hemisphere, and two in the left; for individual electrode coverage see Supplementary Table 3. Two patients were additionally implanted with depth electrodes (28 electrodes in total), which were not analyzed in this study. An additional 35 channels which did not record any signal were also excluded. Recordings were sampled at 1000Hz (CPMC), 3051.76Hz (Stanford, UCSF) or 1535.88Hz (Stanford) and resampled to 1000Hz offline. A high-pass filter was applied online to the signal at either 0.1Hz (five patients, increased to 0.3Hz for parts of the recording in two of the patients), 0.16Hz (one patient) or 0.5Hz (four patients). Electrodes manifesting ictal spikes or persistent noise were identified visually and removed from further analysis (0-38 electrodes per patient, 125 in total; Supplementary Table 3; only analyzed electrodes are plotted in Figure 1a and all other visualizations of electrode positions). Electrodes were re-referenced offline to the average potential of all noise-free electrodes (per patient). Line noise (60Hz and harmonics) was removed offline by a custom made notch filter, designed to remove persistent oscillations (not transients)[52]. All data processing and analysis was done in Matlab (Mathworks, Natick, MA) using custom code, 'pop_eegfiltnew' from EEGLAB toolbox[98], and code from MVPA-Light[99] toolbox for decoding analyses. Colormaps were created using cbrewer[100] with some adaptations.

Electrode localization

Electrodes were localized manually using BioImageSuite (www.bioimagesuite.org) on a post-operative Computed Tomography (CT) scan co-registered to a pre-operative MR scan using the FSL software

package[101]. Individual patient brain images were skull-stripped and segmented using FreeSurfer (http://surfer.nmr.mgh.harvard.edu). Localization errors (resulting from co-registration errors or anatomical mismatch between pre- and post-operative images) were reduced using a custom procedure which jointly minimizes the squared distance between all electrodes within a single electrode array or strip and the cortical pial surface. Individual patients' brains and electrode coordinates were co-registered to a common brain template (FreeSurfer's fsaverage template) using surface-based registration[102], which preserves the mapping of electrode locations to anatomical landmarks and each cortical surface was resampled to a standardized mesh using SUMA[103] (see [9] for more details). Cortical electrodes were assigned to one of six anatomical regions of interest (ROIs) based on the FreeSurfer automatic parcellation (ventral-temporal, occipital, prefrontal, parietal, sensorimotor and lateral-temporal; Figure 1a and Supplementary Table 2). Prefrontal electrodes were further divided into lateral-prefrontal and orbitofrontal cortices, and occipital electrodes were divided into retinotopic and non-retinotopic based on a probabilistic map of visual topographic regions[104] (see electrode locations in Supplementary Figure 1f). Twelve noise-free electrodes located over medial regions (mostly in the precuneus or cingulate cortex) were excluded from further analysis due to their paucity. Visualization of electrode positions was based on surface registration to an MNI152 standard-space T1-weighted average structural template image.

High-frequency activity (HFA) estimation

We focus analysis on high-frequency activity (HFA, 70-150Hz), previously shown to track firing rate in humans[35–37] and other primates[105,106]. We excluded the low-gamma range, which was used in a previous study with this data[9], since it was shown to manifest distinct spectral and functional properties from the HFA signal[107,108]. To estimate the HFA time-course we band-pass filtered the whole signal in eight sub-ranges of 10Hz width between 70 and 150Hz (EEGLAB's FIR Hamming window, using the function pop_eegfiltnew.m[98]). We then extracted the instantaneous amplitude in each band using the Hilbert transform and normalized by dividing the signal by the mean amplitude in that range. Finally, we averaged the amplitude traces from all bands. Normalization was done to account for the 1/f profile of the power spectrum, which results in reduced contribution of the high frequencies relative to the lower frequencies. Trial segments were defined from -300ms to 1600ms around each stimulus onset, and baseline corrected by subtracting the mean HFA signal in the 300ms prior to stimulus onset from the entire trial segment. Trials containing excessive noise were excluded from analysis. The resulting HFB time-courses were smoothed by a 50ms moving window (smoothed time-courses are used in all analyses unless noted otherwise).

In the following sections we describe the four parts of the analysis in detail: Characterization of single electrodes, Characterization of multivariate state-space responses, Decoding of category information, and Characterization of exemplar specific information. Following description of the dependent measures in each section, we describe the approach to statistical analysis. Note that we use a multiverse approach, testing hypotheses in multiple ways to ensure the robustness of the results[109].

Characterization of single electrodes

Visual responsiveness

Visual responsiveness was tested separately for each of the four categories (considering watches and other objects separately), in four non-overlapping "stimulus on" windows: 100-300ms, 300-500ms, 500-700ms and 700-900ms after stimulus onset, considering only trials with durations of 900ms or longer. For each category and each time-window, we compared the mean HFA signal during the "stimulus on" window to the mean HFA signal 200ms prior to stimulus onset (*two-tailed* paired t-test; averaging was done prior to smoothing the HFA trace to avoid information leakage between windows). We used Bonferroni correction across windows, and FDR correction (according to Benjamini-Hochberg procedure[39]) across electrodes, thus electrodes with $q_{FDR} < 0.05/4$ in at least one of the four windows were considered responsive to that category. Note that because the test was two-tailed, electrodes were considered responsive both when activity during the "stimulus on" window was increased relative to the pre-stimulus window, and when it decreased. Changes in response sign between windows or categories were rare (<5% of responsive electrodes), and therefore electrodes were classified as responding positively (increase from baseline; 346/430 electrodes; Figure 1d-g, Supplementary Figure 1a-d, f-g) or responding negatively (decrease from baseline; 84/430 electrodes; Supplementary Figure 1e) based on the sign of the sum of t-statistics from all tests (across all categories and all time-windows). Both types of electrodes were used in all analyses except Figure 1d-g and Supplementary Figures 1a-f, 4 (dark lines), 8b, h, 9d.

Response latencies and attenuation of responses after the onset

To calculate response latencies, we first computed the mean HFA response of each electrode (using only stimuli from categories which the electrode was responsive to, with stimuli duration of 900ms or longer). Peak response magnitude (Figure 1e, Supplementary Figure 1b) and peak response time (Figure 1f, Supplementary Figure 1c) were defined as the maximal HFA value between 0-900ms after onset and the time-point when this was achieved, respectively. Half-peak response time (Supplementary Figure 1g) was defined as the time-point between stimulus onset and peak response time at which the response magnitude was closest to half of the peak response magnitude. Relative attenuation of response magnitude (Figure 1g, Supplementary Figure 1d) was calculated by taking the difference between the peak response and the mean response magnitude 800-900ms after stimulus onset (again in response to stimuli with duration of 900ms or longer), and scaling by the peak response:

$$\text{response attenuation (\% from peak):} \quad 100\,{(P_{mag} - E_{mag})}/{P_{mag}}$$

where P_{mag} is the peak response magnitude and E_{mag} is the mean response 800-900ms after stimulus onset.

Category selectivity

Electrodes were defined as category selective if they showed differential responses to stimuli of the four categories in at least one of the four windows used above for stimuli longer than 900ms (one-way ANOVA, Bonferroni corrected across time-windows and FDR corrected across electrodes, similarly to the responsiveness criterion; Supplementary Figure 1h). For each category selective electrode, we defined a

selectivity time-course by extracting the percent of variance in responses explained by category information on a time-point by time-point basis ($100 \cdot \eta^2$, where η^2 is the effect size measure from a time-point by time-point one-way ANOVA; Figure 1h-k). Comparing each pair of categories separately using a two-tailed t-test and using Cohen's d for a time-point by time-point effect size measure showed similar patterns (Supplementary Figure 1i-j).

Characterization of multivariate state-space responses

The main multivariate analyses (including decoding and exemplar analyses) were performed grouping together responses from multiple patients (c. f. [110–115]). Yet, to ensure no result is driven by a single patient we repeated the central analyses in single patients and found highly consistent results (Supplementary Figures 5d-e, 6a-b, 8d).

Analysis of state-space trajectories

The multivariate neural state at a specific time instance $\vec{s}^t = (s_1^t, s_2^t, \dots, s_E^t)$ is defined as the pattern of recorded activity across electrodes[41–43] (s_e^t is the response of electrode e at time t and E is the number of responsive electrodes in a region). State-space trajectories record the changing multivariate states across time. We quantified the global multivariate response (Figure 2b, d, Supplementary Figures 2b, 3a, c) using the time-point by time-point Euclidean distance of the trajectory from neural state prior to stimulus onset. Since all trials were baseline corrected, this is equivalent to computing the L^2 norm of the vector of responses across electrodes.

$$\text{global multivariate response at time t:} \quad \|\vec{s}^t\| = \sqrt{\sum_{e=1}^{E} (s_e^t)^2}$$

To quantify whether the neural trajectories tracked the duration of the stimulus we also computed the time-point by time-point distance between the responses to stimuli of different durations (Supplementary Figure 3b, d).

$$\text{distance between trajectories at time t:} \quad \|\vec{s}_1{}^t - \vec{s}_2{}^t\|$$

where $\vec{s}_1$ and $\vec{s}_2$ are response trajectories to different stimulus durations.

State transition speed (Figure 2c, Supplementary Figure 2c) was quantified as the distance travelled by the neural state-space trajectory in 1ms (corresponding to our sampling rate; see ref [44] for a similar approach).

$$\text{state transition speed at time t:} \quad \|\vec{s}^t - \vec{s}^{t-1}\|$$

To evaluate the amount of category selectivity in the neural trajectories (Figure 2e, Supplementary Figure 3e) we calculated the dispersion between response trajectories to different categories: First, we computed the square of the Euclidean distance (L^2 norm squared) of each trajectory from the mean response across all categories. We defined the category selectivity index as the square root of the mean of these distances

(analogous to computation of standard deviation in a one-dimensional distribution), and z-scored relative to a permutation null distribution as described in the next section.

$$\text{category selectivity index at time t: } \sqrt{\underset{C}{\text{mean}} \left\| \vec{s}_C^t - \underset{C}{\text{mean}}(\vec{s}_C^t) \right\|^2}$$

where $\vec{s}_C^t$ is the neural response to category C at time t, and $\underset{C}{\text{mean}}$ is the average over all four categories (faces, watches, other objects and animals).

Multivariate response peaks and multivariate response attenuation were calculated similarly to the univariate case.

Statistical testing and confidence intervals

To evaluate statistically the dependence of multivariate response trajectories on stimulus duration (Figure 2d and Supplementary Figure 3a-d) we used permutation testing with max-statistic control for multiple comparisons[116] (N_{perm}=1000, with separate permutations for 900-300 and 1500-900 comparisons). Stimulus duration is expected to influence the response only at the times when one stimulus is still presented, and the other has already ended. Therefore, we considered only the time-points between 300 and 900ms for the first comparison and between 900 and 1500ms for the second (results were highly similar considering the entire time-course). In each permutation we shuffled the duration labels across trials, before averaging the trials of each duration to construct the surrogate duration-specific state-space trajectory. We then computed the relevant statistic (difference between global responses, difference between trajectories) for each time-point. Thus, we created N_{perm} permutation statistic time-series (N_{perm} x N_{time} matrix). Next, we computed the mean and standard deviation across permutations for each time-point (column means and standard deviations) and used these to standardize (z-score) all permutation time-series. To create the surrogate distribution, we now extracted from each standardized permutation time-series the maximal z-value across all time-points. We also z-scored the values of the true unpermuted statistic in the same manner, using the means and standard deviations of the permutation matrix (as under the null hypothesis the true statistic comes from the same distribution as the permutation statistics). Finally, time-points when the unpermuted z-scored statistic was larger than 95% of the null distribution were considered significant (one-sided test). The procedure to test for significant category information (Figure 2e and Supplementary Figure 3e) was similar, except that we shuffled the category affiliation across trials and all time-points were considered in the calculation.

Confidence intervals (Supplementary Figure 2d-e) were computed using a jackknifing approach[117] ($N_{iterations}$=1000, out of $\prod_{p=1}^{10} n_p$ possible combinations per category, where n_p is the number of trials of that category viewed by patient p). In each iteration, for each patient, the state-space response trajectory to each category was estimated by averaging the response to $n_p - 1$ trials (separately for each region). We then merged the response trajectories of single patients (each with its own electrodes) to form a single response trajectory with all electrodes in each cortical ROI. Summary statistics (percent attenuation, speed difference etc.) were computed in the same way as for the neural response calculated based on all images.

Decoding category information

Time-point by time-point classification

To quantify the representation of visual category information in each cortical region, we trained a linear classifier for each time-point to distinguish between trials presenting one category (e.g., faces) vs another (e.g., watches), using the HFA amplitude of all responsive electrodes in the region as features for the classifier[45] (this was done for each pair of categories separately; Figure 3a). For all analyses under this section (Figures 3-5, Supplementary Figures 4-7) the data was downsampled to 200Hz to reduce computation time. To avoid overfitting, decoding analyses were carried out using five-fold cross validation, and to minimize variability stemming from the stochasticity in fold assignment, we repeated the five-fold cross validation procedure five times and averaged the results. In each iteration of the calculation, the category containing more trials was undersampled to balance the number of trials in both categories and we ensured that both categories were represented roughly equally in all folds.

We used regularized Linear Discriminant Analysis (LDA)[118], as implemented in the MVPA-Light Matlab-based toolbox[99] (downloaded on 04/02/2021). LDA attempts to find a one-dimensional projection of the data (a linear weighting of the different electrodes) with maximal separation between the categories. This is done by simultaneously trying to maximize the distance between the mean responses to each category ("signal"), while minimizing the variability in responses to each of the categories ("noise"). Formally, LDA classifier weights (the projection vector) are computed according to the following formula:

$$\text{LDA classifier weights:} \quad \vec{w} = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_2)$$

where $\vec{\mu}_1$ and $\vec{\mu}_2$ are the mean responses to category 1 and category 2 (e.g., faces and watches), and Σ is the pooled covariance matrix (weighted sum of the covariances of each category, in accordance with the number of exemplars in each category). The predicted category for each trial at each time-point ($\vec{s}^t$) is given by comparing $\vec{w}^T\vec{s}^t$ to a set threshold θ:

$$\text{classifier prediction:} \quad \begin{cases} \vec{w}^T\vec{s}^t > \theta \longrightarrow \text{category 1} \\ \vec{w}^T\vec{s}^t < \theta \longrightarrow \text{category 2} \end{cases}$$

The threshold θ can be modified to alter the balance of true positives and false positives (trials from category 1 wrongly classified as category 2 and vice versa), resulting in a full curve characterizing classifier performance. This curve, termed the receiver operating curve (ROC), depicts the false positive rate (FPR) and the true positive rate (TPR) for each threshold θ. We quantified the classifier's performance by using the area under the ROC curve (AUC):

$$\text{area under the curve (AUC):} \quad \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(x)\right) dx$$

When the classifier contains no category information the AUC is equal to 0.5 (50%, chance level performance), and when the classes are fully separable the AUC is equal to 1 (100%, perfect performance).

We used AUC, as opposed to accuracy, since AUC is more robust to unequal class sizes, and is not influenced by changes in the overall magnitude of the response when the pattern across electrodes remains similar.

To reduce the influence of random noise fluctuations on the weight estimate we used a shrinkage estimator for the covariance matrix. Instead of using empirical covariance in the equation above, we use a combination of the empirical covariance (Σ) and the identity matrix (I), weighted according to the regularization parameter λ:

$$\text{shrinkage regularization of the covariance matrix:} \quad \hat{\Sigma} = (1-\lambda)\Sigma + \lambda \frac{trace(\Sigma)}{p} \mathrm{I}$$

where p is the number of features of the classifier (equal to the number of responsive electrodes). Multiplying the identity matrix by the mean value on the diagonal ensures that the trace of the covariance is preserved, which helps to mitigate the bias introduced by this step. The regularization parameter λ was estimated using the Ledoit-Wolf formula[119] (default implementation in MVPA-Light[99], for more explanation on the rationale of shrinkage regularization see [120]).

The majority of decoding analyses were carried out on responses to 92 unique images which were seen by all patients for 900ms or longer (faces, 30; watches, 32; other objects 18; animals 12; Figures 3b-e, 4a-e, Supplementary Figures 4, 5a-c, 6a-e). To avoid overfitting[45], repetitions of the same image were averaged to ensure the same image is not used simultaneously in the training and testing sets (number of repetitions mean ± std across patients and unique exemplars: 1.5 ± 0.6, range: 1-4, similar for all categories). Using only the first repetition of each image elicited highly similar results. Analyses performed on data from single patients used all unique exemplars viewed by each patient for 900ms or longer (mean ± SEM unique exemplars across patients: faces, 62.5 ± 4.2; watches, 65.5 ± 4.5; objects (non-watch) 39 ± 2.3; animals, 22 ± 1.1; Figures 3f, 4h, Supplementary Figures 5d-e, 6a-b, e). Comparison of bipolar and average montages was performed on patients S8 and S10 together (faces, 54; watches, 58; other objects, 34; animals, 20; Figure 4f-g, Supplementary Figure 6d). To run analyses for each stimulus duration separately (Figure 5, Supplementary Figure 7) we split the patients into two groups (S1-S3, S4-S10). This was done in order to increase the number of exemplars, as due to the randomization the full group shared on average only 4.8 exemplars per category per duration. The number of exemplars shared in each category and duration by the full patient group and each patient sub-group are shown in Supplementary Table 6.

Temporal generalization analysis

To assess whether coding of category information was stable in time we used the temporal generalization method, developed in ref [49] (Figures 3d, 4b, 5c-d, Supplementary Figures 5, 7f). In this method linear classifiers are trained on data from each time-point separately, but tested on data from all time-points, not only the one used for the training data. The result is a matrix of decoding values, with the y-axis indicating the training time point and the x-axis indicating the testing time-point. Successful decoding between time-

points indicates that the pattern separating the categories in one time-point also separates the categories in the other time-point, implying a stable coding of category information in time.

Statistical testing and correction for multiple comparisons

All statistical testing was based on permutation tests (N_{perm}=1000). To construct the null permutation distribution, in each permutation we permuted the category labels of all trials and trained classifiers for all time-points based on the permuted labels (we used the same set of permuted labels to preserve the temporal structure of the data, required for some of the approaches we used to control for multiple comparisons). Peak decoding performance (Figure 3b) was statistically tested by comparing to the peak AUC of each permutation (maximum across time-points), equivalent to a max-statistic approach to control for multiple comparisons (one-sided)[116]. Time-courses and temporal generalization matrices were tested either by using cluster-based permutations to control for multiple comparisons[47] (details below; Figures 3c, 4a, d, f, 5, Supplementary Figures 4, 5c-e, 6a-d, 7b-f) or, when inferences about specific time-points were required, using point-by-point comparisons controlling the false discovery rate (FDR; Benjamini-Hochberg procedure[39], $q_{FDR} < 0.05$, one-sided; this was required since cluster-based permutations are insufficient to establish the latency or temporal extent of effects[48]; Figures 3c-d, 4a-b, Supplementary Figure 5).

Cluster permutation details: First, we selected all time-points with AUC greater or equal to 60% (first-level threshold; changing this threshold led to similar results), then clustered the samples based on temporal adjacency (including both train and test times when applied for temporal generalization matrices) and extracted the sum of AUC values in each cluster (cluster statistic). To construct the null distribution, we performed the same procedure on each of the permutations, but only the maximal cluster value was retained in the null distribution. Clusters from the original (non-permuted) calculation that exceeded 95% of this distribution were considered significant (one-sided test).

Statistical comparison of decoding of stimuli in different durations was performed in a similar way, except the cluster first-level threshold was set to AUC difference of 10% between duration conditions (Figure 5, Supplementary Figure 7). We permuted the category labels and trained classifiers for each duration separately and formed the null distribution by subtracting the decoding results for the short duration (300 or 900ms) from the long duration (900 or 1500ms). For the decoding time-courses, we considered only time-points after the offset of the short stimulus, to reduce the number of simultaneous comparisons and as we predicted a difference only when one stimulus was presented, and the other was not (results were nearly identical without this constraint). For generalization matrices we considered the full matrix since we were also interested in generalization between the onset and the time-period after the offset of the short stimulus.

Characterization of exemplar specific information

Quantifying exemplar representation

To quantify exemplar-level information we used representational similarity analysis (RSA)[54,55], which describes neural representation in terms of the relation between neural responses, that is, the geometry that the responses define. The representational geometry is fully captured by the set of all pairwise dissimilarities between the pattern of responses across electrodes to each pair of stimuli, grouped together in a

representational dissimilarity matrix (RDM). We used two dissimilarity metrics – correlation dissimilarity (Figures 6-8, Supplementary Figures 8b-f, h-i, 9a-b, d-e, 10, 11) and Euclidean dissimilarity (Figure 6, Supplementary Figures 8b-c, g, 9c):

$$\text{correlation dissimilarity:} \quad 1 - r(\vec{s}_A, \vec{s}_B)$$

$$\text{Euclidean dissimilarity:} \quad \|\vec{s}_A - \vec{s}_B\|$$

where $\vec{s}_A$ and $\vec{s}_B$ are the vector of HFA responses across electrodes to stimulus A and stimulus B (in a specific time-point), $r(\vec{u}, \vec{v})$ is the pearson correlation of $\vec{u}$ and $\vec{v}$ and $\|\cdot\|$ is L_2 norm.

To avoid overfitting to incidental noise fluctuations, exemplar-level analyses were only conducted on stimuli which were viewed at least twice (akin to a cross-validation procedure), for 900ms or longer (to focus on the sustained response). Four patients viewed only 27 exemplars from the analyzed categories twice, and one patient did not view any exemplars twice, thus, to ensure sufficient sampling of all categories we performed the main analyses in this section on a group of five patients (194 responsive electrodes) who all viewed the same set of 60 unique images at least twice (faces, 18; watches, 19; other objects, 13; animals, 10; Figures 6-8, Supplementary Figures 8b, e-h, 9a-d, 10-11). Results from a larger eight patient group (i.e., excluding one patient who did not view any exemplars twice and one which had excessive noise in many trials; 331 responsive electrodes) with only 18 unique exemplars are presented as a control in Supplementary Figures 8c, i and 9e (faces, 6; watches, 7; other objects, 2; animals, 3). Electrode locations of both patient groups are shown in Supplementary Figure 8a. Analyses performed on data from single patients used all unique exemplars viewed by each patient for 900ms or longer (Supplementary Figure 8d). In cases of more than two repetitions, we averaged responses to form two independent responses.

We computed the dissimilarity structure separately for each repetition and time-point and compared between repetitions (reliability analysis) and between time-points (stability analysis) using Spearman correlation, which is more robust to changes in the overall magnitude of activity relative to Pearson correlation. As with the decoding analysis, for all analyses under this section (Figures 6-8, Supplementary Figures 8-11) the data was downsampled to 200Hz to reduce computation time.

Representation reliability across repetitions

Item reliability index (Figures 6a, c-e, 7a-d, 8, Supplementary Figures 8b-i, 9a, c-e, 10b-c, 11): If exemplars are represented reliably, the location of each exemplar within the neural geometry should remain stable. Thus, for each stimulus presentation (Figure 6a, red star), we computed the correlation of the vector of dissimilarities to all other images in repetition 1 (filled shapes) and the vector of dissimilarities to all other images of repetition 2 (empty shapes). We then averaged the obtained correlations across all images and repetitions, and z-scored the result using a permutation null distribution (N_{perm}=1000). To construct the null distribution, for each image presentation we shuffled the stimulus identity of all images from the *other* repetition and repeated the process (i.e., if the presented image was from repetition 1 (as in Figure 6a) we shuffled repetition 2 and if the presented image was from repetition 2 we shuffled repetition 1).

Geometry reliability index (Figures 6b-c, f-g, 7e, Supplementary Figures 8e-i, 9, 10d-e): We designed a second measure of reliability to capture global aspects of the representational geometry, by comparing the full dissimilarity structure across repetitions. We first computed the RDM between all stimulus presentations from both repetitions (Figure 6b, top), containing four distinct sets of dissimilarities: within repetition 1 (blue), within repetition 2 (green) and two between repetition 1 and repetition 2 (yellow and red). Second, we paired each within repetition dissimilarity set with one between repetitions dissimilarity set and averaged the dissimilarities in each pair (resulting in Geometry 1 and Geometry 2 in Figure 6b; averaging different pairs of submatrices of Figure 6b or correlating each pair of geometries separately and then averaging the results led to similar results). Considering both within repetition and between repetition dissimilarities ensures not only preservation of the geometry across repetitions, but also a similar state-space location between repetitions. Third, we unfolded Geometry 1 and Geometry 2 into vectors and correlated between them (exemplified in Figure 6c bottom). Finally, we z-scored the result using a permutation null distribution (N_{perm}=1000). We constructed the null distribution by shuffling the identity of all *single exemplars* in repetition 2 and repeating the procedure. Importantly, in cases where some exemplars are represented similarly within the geometry (indistinguishable representations) this approach is likely to result in insignificant Geometry reliability even if the geometry is fully maintained between repetitions. Therefore, this is a test of both representational reliability between repetitions and of discriminability between exemplars.

Representation stability in time

We tested for temporal stability of the representation by comparing representational structures across time-points (Figures 7c-e, 8d, Supplementary Figures 9, 10c, e, 11e). This was done similarly to the reliability analyses; only representational structures were computed in different time-points. Specifically, to calculate Item stability between t_1 and t_2 we computed for each stimulus the neural dissimilarity between the representation of the stimulus at time t_1 and representation of all other exemplars in the *same* repetition also at t_1 (y-axis of generalization plots) and the neural dissimilarity between the stimulus representation at time t_2 and representation of all other exemplars in the *other* repetition (x-axis of generalization plots). To calculate Geometry stability between t_1 and t_2, we correlated Geometry 1 in t_1 (y-axis) with Geometry 2 in t_2 (x-axis). When $t_1 = t_2$ (diagonal of stability matrices) this is equivalent to calculation of Item and Geometry reliability, respectively. Note that by correlating dissimilarity structures at different repetitions (for at least one of the exemplars), we avoid overfitting to spontaneous fluctuations unrelated to stimulus processing, which are known to exhibit many temporal dependencies[57–59].

Controlling for category information

To test whether exemplar-level information is present beyond category differences we designed four models of category information (Figure 8a, see description below) and tested whether exemplar-level information is reliable and stable after removing category information from the neural RDM using partial correlation (i.e., partialling out the model RDMs; Figure 8a-d, Supplementary Figure 10). For Item reliability\ temporal stability we used a single row of the model RDM for each stimulus (treating the RDM as a symmetric matrix, but excluding the diagonal), and for Geometry reliability\ temporal stability we partialled out the full model RDM unfolded into a vector.

The four category models were: (1) "Single-category" model – exemplars within each category are fully similar to each other and between categories exemplars are fully dissimilar to each other (no hierarchy between categories). The other three models group individual categories into high level groups; exemplars within categories are assumed fully similar to each other, exemplars between high-order groups are fully dissimilar, and exemplars in different categories within each high-level group are 50% similar. (2) "Low-level" model – relation between categories based on low-level visual similarity. This model groups watches and faces (round photos) in one high-order category, and (non-watch) objects and animals (pointy illustrations) in another high-other category. (3) "Sematic" model – relation between categories based on semantic similarity (animacy). This model groups faces and animals in one high-order category and watches and other objects in another high-order category. (4) "Face-vs-rest" model – in this model the face category is distinct from all other categories which are grouped together to one high-order category. We constructed this model due to the unique social importance of faces and the known specialization in face representation[121,122] (also supported by our results, Figure 2a).

Statistical testing and correction for multiple comparisons

All statistical testing was permutation-based (N_{perm}=1000). Construction of the null distribution is detailed above (section 'Representation reliability across repetitions', different procedures for Item reliability\ stability and Geometry reliability\ stability). In each permutation, we used the same set of permuted labels for all time-points to preserve temporal properties of the data. Z-scoring was performed in each time-point separately, for both the original (non-permuted) score and the permutation results. Statistical testing and control for multiple comparisons was done similarly to the decoding analyses (all one-sided): peak reliability indices (Figures 6d-g, 8e, Supplementary Figures 8b-d, 11a-c) were tested by comparing to the peak of each permutation (max-statistic approach[116]. Reliability time-courses were tested using cluster-based permutations[47] (sum of reliability indices as the cluster-statistic). First-level threshold was set to z = 1.5 SD for the main analyses (Figures 7a-b, 8c, Supplementary Figures 8e-i, 10b, d) and z = 0.75 SD for the single category analyses (adjusted to accommodate the larger temporal smoothing window in these analyses; Figure 8g, Supplementary Figure 11d-e), in both cases similar results were obtained with other thresholds. The time-courses of correlation between each category RDM model and the neural RDM (Supplementary Figure 10a) were also corrected using cluster-based permutations, with the sum of the correlations as the cluster-statistic and ρ=0.0391 as the first-level threshold (corresponding to a p-value of 0.05). Stability (generalization) matrices were tested using point-by-point comparisons, controlling FDR using Benjamini-Hochberg procedure[39] ($q_{FDR} < 0.05$; Figures 7c, 8d, Supplementary Figures 9, 10c, e).

Acknowledgements

The authors wish to thank members of the Human Cognitive Neuroscience Laboratory (HCNL), as well A. N. Landau and the Landau lab, and R. Malach and R. Broday-Dvir for helpful discussions, and H. T. Vishne for graphics design and multiple late nights. We also thank patients who participated in the experiment and J. Parvizi, R. A. Kuperman, K. I. Auguste and E. F. Chang for enabling data collection from their patients. The work was supported by grant 2013070 from the US-Israel Binational Science Foundation to LYD and RTK. LYD is supported by the Jack H. Skirball research fund. GV is supported by the Azrieli foundation

graduate fellowship. This work is dedicated to the memory of Mrs. Lily Safra, a great supporter of brain research.

Author contributions

LYD and EMG initiated and designed the experiment. GV and LYD conceptualized the analysis. GV designed novel methodology and analyzed the data. RTK supervised data collection, detected pathological activity. EMG took part in data collection, localized electrodes and assisted in preprocessing the data. GV and LYD wrote the paper, RTK reviewed it. LYD and RTK provided funding. LYD supervised the study.

References

1. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).

2. Grill-Spector, K. & Kanwisher, N. Visual recognition: as soon as you know it is there, you know what it is. *Psychol. Sci.* **16**, 152–160 (2005).

3. Land, M. F. & Hayhoe, M. In what ways do eye movements contribute to everyday activities? *Vision Res.* **41**, 3559–3565 (2001).

4. Land, M. F. Eye movements and the control of actions in everyday life. *Prog. Retin. Eye Res.* **25**, 296–324 (2006).

5. Reddy, L. & Kanwisher, N. Coding of visual objects in the ventral stream. *Curr. Opin. Neurobiol.* **16**, 408–414 (2006).

6. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

7. Grill-Spector, K. & Weiner, K. S. The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* **15**, 536–548 (2014).

8. Podvalny, E. *et al.* Invariant temporal dynamics underlie perceptual stability in human visual cortex. *Curr. Biol.* **27**, 155–165 (2017).

9. Gerber, E. M., Golan, T., Knight, R. T. & Deouell, L. Y. Cortical representation of persistent visual stimuli. *Neuroimage* **161**, 67–79 (2017).

10. Crick, F. & Koch, C. Towards a neurobiological theory of consciousness. *seminars in The Neurosciences* **2**, 263–275 (1990).

11. Crick, F. & Koch, C. A framework for consciousness. *Nat. Neurosci.* **6**, 119–126 (2003).

12. Melloni, L., Mudrik, L., Pitts, M. & Koch, C. Making the hard problem of consciousness easier. *Science* **372**, 911–912 (2021).

13. Seth, A. K. & Bayne, T. Theories of consciousness. *Nat. Rev. Neurosci.* **23**, 439–452 (2022).

14. Moutard, C., Dehaene, S. & Malach, R. Spontaneous Fluctuations and Non-linear Ignitions: Two Dynamic Faces of Cortical Recurrent Loops. *Neuron* **88**, 194–206 (2015).

15. Lamme, V. A. F. Towards a true neural stance on consciousness. *Trends Cogn Sci (Regul Ed)* **10**, 494–501 (2006).

16. Fisch, L. *et al.* Neural "ignition": enhanced activation linked to perceptual awareness in human ventral stream visual cortex. *Neuron* **64**, 562–574 (2009).

17. Dehaene, S. & Naccache, L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition* **79**, 1–37 (2001).

18. Mashour, G. A., Roelfsema, P., Changeux, J.-P. & Dehaene, S. Conscious processing and the global neuronal workspace hypothesis. *Neuron* **105**, 776–798 (2020).

19. Malach, R. Local neuronal relational structures underlying the contents of human conscious experience. *Neurosci. Conscious.* **2021**, niab028 (2021).

20. Tononi, G., Boly, M., Massimini, M. & Koch, C. Integrated information theory: from consciousness to its physical substrate. *Nat. Rev. Neurosci.* **17**, 450–461 (2016).

21. Lau, H. & Rosenthal, D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci (Regul Ed)* **15**, 365–373 (2011).

22. Brown, R., Lau, H. & LeDoux, J. E. Understanding the Higher-Order Approach to Consciousness. *Trends Cogn Sci (Regul Ed)* **23**, 754–768 (2019).

23. Boly, M. *et al.* Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? clinical and neuroimaging evidence. *J. Neurosci.* **37**, 9603–9613 (2017).

24. Grasso, M., Haun, A. M. & Tononi, G. Of maps and grids. *Neurosci. Conscious.* **2021**, niab022 (2021).

25. Dehaene, S. *et al.* Cerebral mechanisms of word masking and unconscious repetition priming. *Nat. Neurosci.* **4**, 752–758 (2001).

26. Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends Cogn Sci (Regul Ed)* **10**, 204–211 (2006).

27. Odegaard, B., Knight, R. T. & Lau, H. Should a few null findings falsify prefrontal theories of conscious perception? *J. Neurosci.* **37**, 9593–9602 (2017).

28. Frässle, S., Sommer, J., Jansen, A., Naber, M. & Einhäuser, W. Binocular rivalry: frontal activity relates to introspection and action but not to perception. *J. Neurosci.* **34**, 1738–1747 (2014).

29. Tsuchiya, N., Wilke, M., Frässle, S. & Lamme, V. A. F. No-Report Paradigms: Extracting the True Neural Correlates of Consciousness. *Trends Cogn Sci (Regul Ed)* **19**, 757–770 (2015).

30. Mudrik, L. *et al.* Preregistration file - v2 - May 2020. (2021). Available at: osf.io/f87xq.

31. Parvizi, J. & Kastner, S. Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.* **21**, 474–483 (2018).

32. Stigliani, A., Jeska, B. & Grill-Spector, K. Encoding model of temporal processing in human visual cortex. *Proc Natl Acad Sci USA* **114**, E11047–E11056 (2017).

33. Daniel, S. & Bentin, S. Age-related changes in processing faces from detection to identification: ERP evidence. *Neurobiol. Aging* **33**, 206.e1–28 (2012).

34. Meinhardt-Injac, B. The context congruency effect is face specific. *Acta Psychol (Amst)* **142**, 265–272 (2013).

35. Mukamel, R. *et al.* Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. *Science* **309**, 951–954 (2005).

36. Nir, Y. *et al.* Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. *Curr. Biol.* **17**, 1275–1285 (2007).

37. Manning, J. R., Jacobs, J., Fried, I. & Kahana, M. J. Broadband shifts in local field potential power spectra are correlated with single-neuron spiking in humans. *J. Neurosci.* **29**, 13613–13620 (2009).

38. Leszczyński, M. *et al.* Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. *Sci. Adv.* **6**, eabb0977 (2020).

39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).

40. Bechtold, B *et al.* bastibe/Violinplot-Matlab: A Good Starting Point (v0.1). *Zenodo* (2021) doi: 10.5281/zenodo.4559847.

41. Yuste, R. From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* **16**, 487–497 (2015).

42. Saxena, S. & Cunningham, J. P. Towards the neural population doctrine. *Curr. Opin. Neurobiol.* **55**, 103–111 (2019).

43. Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055–3068 (2021).

44. Weber, J. *et al.* Population coding and oscillatory subspace synchronization integrate context into actions. *BioRxiv* (2021) doi: 10.1101/2021.12.17.473118.

45. Grootswagers, T., Wardle, S. G. & Carlson, T. A. 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 (2017).

46. Carlson, T. A., Grootswagers, T. & Robinson, A. K. An introduction to time-resolved decoding analysis for M/EEG. (2019).

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

48. Sassenhagen, J. & Draschkow, D. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology* **56**, e13335 (2019).

49. King, J. R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn Sci (Regul Ed)* **18**, 203–210 (2014).

50. Jerbi, K. *et al.* Saccade related gamma-band activity in intracerebral EEG: dissociating neural from ocular muscle activity. *Brain Topogr.* **22**, 18–23 (2009).

51. Kovach, C. K. *et al.* Manifestation of ocular-muscle EMG contamination in human intracranial recordings. *Neuroimage* **54**, 213–233 (2011).

52. Keren, A. S., Yuval-Greenberg, S. & Deouell, L. Y. Saccadic spike potentials in gamma-band EEG: characterization, detection and suppression. *Neuroimage* **49**, 2248–2263 (2010).

53. Aru, J., Bachmann, T., Singer, W. & Melloni, L. Distilling the neural correlates of consciousness. *Neurosci. Biobehav. Rev.* **36**, 737–746 (2012).

54. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci (Regul Ed)* **17**, 401–412 (2013).

55. Kriegeskorte, N. & Wei, X.-X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).

56. Walther, A. *et al.* Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage* **137**, 188–200 (2016).

57. Arieli, A., Sterkin, A., Grinvald, A. & Aertsen, A. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science* **273**, 1868–1871 (1996).

58. Nir, Y. *et al.* Interhemispheric correlations of slow spontaneous neuronal fluctuations revealed in human sensory cortex. *Nat. Neurosci.* **11**, 1100–1108 (2008).

59. He, B. J., Snyder, A. Z., Zempel, J. M., Smyth, M. D. & Raichle, M. E. Electrophysiological correlates of the brain's intrinsic large-scale functional architecture. *Proc Natl Acad Sci USA* **105**, 16039–16044 (2008).

60. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).

61. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).

62. Grootswagers, T., Robinson, A. K. & Carlson, T. A. The representational dynamics of visual objects in rapid serial visual processing streams. *Neuroimage* **188**, 668–679 (2019).

63. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).

64. Holy, T. *Generate maximally perceptually-distinct colors.* (MATLAB Central File Exchange, 2011).

65. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* (2008).

66. Baars, B. J. *A cognitive theory of consciousness.* (Cambridge University Press, 1989).

67. Breitmeyer, B. G. Psychophysical "blinding" methods reveal a functional hierarchy of unconscious visual processing. *Conscious. Cogn.* **35**, 234–250 (2015).

68. Mudrik, L. & Deouell, L. Y. Neuroscientific evidence for processing without awareness. *Annu. Rev. Neurosci.* **45**, 403–423 (2022).

69. Spector Shirtz, A. Pain Begins. *Theoria* (2022) doi: 10.1111/theo.12391.

70. Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).

71. Haxby, J. V., Connolly, A. C. & Guntupalli, J. S. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37**, 435–456 (2014).

72. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* **102**, 249-259.e4 (2019).

73. Kohn, A. *et al.* Principles of corticocortical communication: proposed schemes and design considerations. *Trends Neurosci.* **43**, 725–737 (2020).

74. Fink, S. B., Kob, L. & Lyre, H. A structural constraint on neural correlates of consciousness. *PhiMiSci* **2**, (2021).

75. Lau, H., Michel, M., LeDoux, J. E. & Fleming, S. M. The mnemonic basis of subjective experience. *Nat. Rev. Psychol.* (2022) doi: 10.1038/s44159-022-00068-6.

76. Bickle, J. Multiple Realizability (Stanford Encyclopedia of Philosophy). (1998).

77. VanRullen, R. & Koch, C. Is perception discrete or continuous? *Trends Cogn Sci (Regul Ed)* **7**, 207–213 (2003).

78. Herzog, M. H., Drissi-Daoudi, L. & Doerig, A. All in Good Time: Long-Lasting Postdictive Effects Reveal Discrete Perception. *Trends Cogn Sci (Regul Ed)* **24**, 826–837 (2020).

79. Block, N. Two neural correlates of consciousness. *Trends Cogn Sci (Regul Ed)* **9**, 46–52 (2005).

80. Lamme, V. A. F. How neuroscience will change our view on consciousness. *Cogn. Neurosci.* **1**, 204–220 (2010).

81. Pitts, M. A., Metzler, S. & Hillyard, S. A. Isolating neural correlates of conscious perception from neural correlates of reporting one's perception. *Front. Psychol.* **5**, 1078 (2014).

82. Bellet, J. *et al.* Decoding rapidly presented visual stimuli from prefrontal ensembles without report nor post-perceptual processing. *Neurosci. Conscious.* **2022**, niac005 (2022).

83. Kapoor, V. *et al.* Decoding internally generated transitions of conscious contents in the prefrontal cortex without subjective reports. *Nat. Commun.* **13**, 1535 (2022).

84. Szczepanski, S. M. & Knight, R. T. Insights into human behavior from lesions to the prefrontal cortex. *Neuron* **83**, 1002–1018 (2014).

85. Carlén, M. What constitutes the prefrontal cortex? *Science* **358**, 478–482 (2017).

86. Bar, M. *et al.* Top-down facilitation of visual recognition. *Proc Natl Acad Sci USA* **103**, 449–454 (2006).

87. Trapp, S. & Bar, M. Prediction, context, and competition in visual recognition. *Ann. N. Y. Acad. Sci.* **1339**, 190–198 (2015).

88. Yuval-Greenberg, S., Tomer, O., Keren, A. S., Nelken, I. & Deouell, L. Y. Transient induced gamma-band response in EEG as a manifestation of miniature saccades. *Neuron* **58**, 429–441 (2008).

89. Lumer, E. D., Friston, K. J. & Rees, G. Neural correlates of perceptual rivalry in the human brain. *Science* **280**, 1930–1934 (1998).

90. Wilterson, A. I., Nastase, S. A., Bio, B. J., Guterstam, A. & Graziano, M. S. A. Attention, awareness, and the right temporoparietal junction. *Proc Natl Acad Sci USA* **118**, (2021).

91. Koch, C., Massimini, M., Boly, M. & Tononi, G. Neural correlates of consciousness: progress and problems. *Nat. Rev. Neurosci.* **17**, 307–321 (2016).

92. Haun & Tononi. Why does space feel the way it does? towards a principled account of spatial experience. *Entropy* **21**, 1160 (2019).

93. Culham, J. C. & Kanwisher, N. G. Neuroimaging of cognitive functions in human parietal cortex. *Curr. Opin. Neurobiol.* **11**, 157–163 (2001).

94. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).

95. Rigotti, M. *et al.* The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).

96. Ringo, J. L. Stimulus specific adaptation in inferior temporal and medial temporal cortex of the monkey. *Behav. Brain Res.* **76**, 191–197 (1996).

97. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci (Regul Ed)* **10**, 14–23 (2006).

98. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).

99. Treder, M. S. MVPA-Light: A Classification and Regression Toolbox for Multi-Dimensional Data. *Front. Neurosci.* **14**, 289 (2020).

100. Charles. *cbrewer: colorbrewer schemes for Matlab.* (MATLAB Central File Exchange, 2015).

101. Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W. & Smith, S. M. FSL. *Neuroimage* **62**, 782–790 (2012).

102. Fischl, B., Sereno, M. I., Tootell, R. B. & Dale, A. M. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum. Brain Mapp.* **8**, 272–284 (1999).

103. Argall, B. D., Saad, Z. S. & Beauchamp, M. S. Simplified intersubject averaging on the cortical surface using SUMA. *Hum. Brain Mapp.* **27**, 14–27 (2006).

104. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic maps of visual topography in human cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).

105. Rasch, M. J., Gretton, A., Murayama, Y., Maass, W. & Logothetis, N. K. Inferring spike trains from local field potentials. *J. Neurophysiol.* **99**, 1461–1476 (2008).

106. Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J. & Hsiao, S. S. Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008).

107. Ray, S. & Maunsell, J. H. R. Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biol.* **9**, e1000610 (2011).

108. Bartoli, E. *et al.* Functionally distinct gamma range activity revealed by stimulus tuning in human visual cortex. *Curr. Biol.* **29**, 3345-3358.e7 (2019).

109. Steegen, S., Tuerlinckx, F., Gelman, A. & Vanpaemel, W. Increasing transparency through a multiverse analysis. *Perspect. Psychol. Sci.* **11**, 702–712 (2016).

110. Mesgarani, N., Cheung, C., Johnson, K. & Chang, E. F. Phonetic feature encoding in human superior temporal gyrus. *Science* **343**, 1006–1010 (2014).

111. Hamilton, L. S., Edwards, E. & Chang, E. F. A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. *Curr. Biol.* **28**, 1860-1871.e4 (2018).

112. Grossman, S. *et al.* Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).

113. Norman, Y. *et al.* Hippocampal sharp-wave ripples linked to visual episodic recollection in humans. *Science* **365**, (2019).

114. Goldstein, A. *et al.* Shared computational principles for language processing in humans and deep language models. *Nat. Neurosci.* **25**, 369–380 (2022).

115. Norman-Haignere, S. V. *et al.* A neural population selective for song in human auditory cortex. *Curr. Biol.* **32**, 1470-1484.e12 (2022).

116. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).

117. Miller, R. G. The jackknife-a review. *Biometrika* **61**, 1–15 (1974).

118. Friedman, J. H. Regularized Discriminant Analysis. *J. Am. Stat. Assoc.* **84**, 165–175 (1989).

119. Ledoit, O. & Wolf, M. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **88**, 365–411 (2004).

120. Blankertz, B., Lemm, S., Treder, M., Haufe, S. & Müller, K.-R. Single-trial analysis and classification of ERP components--a tutorial. *Neuroimage* **56**, 814–825 (2011).

121. Bentin, S., Allison, T., Puce, A., Perez, E. & McCarthy, G. Electrophysiological studies of face perception in humans. *J. Cogn. Neurosci.* **8**, 551–565 (1996).

122. Kanwisher, N. & Yovel, G. The fusiform face area: a cortical region specialized for the perception of faces. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2109–2128 (2006).