Reproducibility Distinguishes Conscious From Non-conscious Neural Representations

Aaron Schurger[1,2*], Francisco Pereira[1,2], Anne Treisman[1], and Jonathan D. Cohen[1,2]

[1] Department of Psychology, Princeton University, Princeton, New Jersey, USA

[2] Center for the Study of Brain, Mind, and Behavior, Princeton University, Princeton, New Jersey, USA

* To whom correspondence should be addressed. E-mail:

schurger@princeton.edu

One-sentence summary:

A Euclidean approach to the analysis of fMRI data reveals that category-specific neural activation patterns are significantly more reproducible for seen versus unseen objects.

Abstract

What qualifies a neural representation for a role in subjective experience? Previous evidence suggests that the duration and intensity of the neural response to a sensory stimulus are factors. We introduce another attribute – the reproducibility of a pattern of neural activity across different episodes – that predicts specific and measurable differences between conscious and non-conscious neural representations indepedently of duration and intensity. We found that conscious neural activation patterns are relatively reproducible when compared to non-conscious neural activation patterns corresponding to the same perceptual content. This is not adequately explained by a difference in signal-to-noise ratio.

Though once controversial, it is now widely accepted that sensory-perceptual information can be processed by the brain, even at the semantic level, without that information "reaching" or "entering" awareness (*1-3*). But what does it mean for neural information to "reach" awareness? Once the information has been encoded in neural activity, what else has to happen for it to become part of one's subjective reality? A growing body of evidence suggests that the intensity of activation in areas that encode the contents of perception (e.g. ventral-temporal cortex) is one determinant of whether or not that information contributes directly to subjective experience (*4-7*). However, local enhancement of a cortical sensory signal is also associated with attention (*8*), which can be independent of awareness (*9-11*). Therefore, there may be additional features other than the intensity of neural activity that distinguish conscious from non-conscious neural information.

Kinsbourne (*12*) proposes three interacting properties that collectively determine whether or not a neural representation will contribute directly to subjective experience: (a) the duration and (b) the intensity of a pattern of activity, and (c) the *coherence* of that pattern of activity with the dominant "configuration" of neural activity at the global level. Here we propose that another attribute of neural activity patterns — reproducibility — characterizes conscious representations. We define reproducibility as the similarity of patterns of neural activity across different instances of the same percept. We focus specifically on reproducibility because it is measurable, and therefore empirically testable. A corollary of our proposal that conscious representations are more reproducible is that unconscious representations are more variable, even as they may carry information within a given episode.

We used functional magnetic resonance imaging (fMRI) to measure brain activity while subjects performed a simple visual category-discrimination task (N = 12 subjects (*13*)). Stimuli were simple line drawings of faces and houses (12 of each), rendered in two opposing but isoluminant colors (Fig. 1 and SOM). Visibility of the stimuli was manipulated using dichoptic color masking (DCM; (*7*) and Fig. 1). Subjects were asked to identify the category of the stimulus (face or house) on each trial, guessing if necessary, and also to wager ("high" or "low", for monetary rewards) on the accuracy of each of their perceptual decisions (*14 – 16*). Wagering was used as a collateral index of subjects' awareness of the object.

For visible stimuli, performance was at or near 100% correct for all 12 subjects and all wagers were "high". For invisible stimuli, task performance was only marginally different from chance (54 +/- $2.5_{[SEM]}$ % correct; $p < 0.06$, one-tailed), and sensitivity of high wagers to correct responses (wagering d-prime, or d'; see SOM) was not different from zero (mean d' = 0.015 +/- $0.11_{[SEM]}$; $p = 0.45$, one-tailed). For invisible stimuli, wagering d' and overall willingness to place high wagers were not significantly correlated across subjects ($r = 0.33$, $p > 0.30$, $n = 12$). This reassures against the possibility that wagering d' was artificially low due to an interaction with a wagering bias (*16*). The proportion of high wagers (for invisible stimuli) was similar for faces and houses (0.20 and 0.19, respectively).

Subjects were always aware of a visual event - a yellowish flickering square - and this provoked significant activation in and of itself. What varied was subjects' awareness of an object embedded in the square. We used multivariate pattern analysis to ascertain how the encoding of perceptual information differs depending upon whether or not that information is present in

subjective experience (*17*). Thus, in our analyses we focused specifically on the patterns of activation corresponding to the perceptual information of which the subject was or was not aware – the category of the object.

To verify the neural representation of category-specific information for both visible and invisible stimuli, we attempted to discriminate the category of the stimulus (faces versus houses) based on the spatial pattern of neural activity in the temporal lobes (derived statistically from each run of functional data (*13*)). We did this independently for the visible and invisible stimuli, using a Gaussian Naïve Bayes classifier (*18*). We focused our analyses on the temporal lobes, because these are widely viewed as being critical for high level perceptual representation of visual information (*19*). Mean accuracy of the classifier (% correct averaged across 12 subjects) was significantly different from chance (50%) for both visible (63% correct; $t = 3.82$, $p < 0.002$) and invisible (58% correct; $t = 2.53$, $p < 0.02$) stimuli (see table 1). The difference in accuracy for visible versus invisible stimuli was not significant ($p < 0.2$, one-tailed paired-samples t-test). It might be expected that as long as the classifier performed above chance on both types of stimuli then it would also perform well when trained on one type and tested on the other (*20*). However, this was not the case for these stimuli (table 1).

Each round of training/testing of the classifier involved a dimensionality-reduction step, wherein we determined which voxels (features) varied most consistently as a function of stimulus category (feature selection) *separately for visible and invisible stimuli* (*13*). Training / testing of the classifier was then performed on these smaller feature spaces ("selections"). Our approach involved examining the patterns of activity within these selections of voxels, on the assumption

that these would reveal properties of information encoding under conditions of conscious and non-conscious perception.

Activation patterns

Treating patterns of activation as vectors allows us to test hypotheses about the properties of neural information, independently of specific loci and their level of activity. The angle between two activation vectors reflects differences in the contents of perception, while the norm of each vector corresponds to the intensity of the information being encoded. We can then define reproducibility as *the similarity in the pattern of activity across different instances of the same stimulus category, among voxels that carry relevant information*. This can be measured by computing the trial-to-trial variability of the vector angle in the space of the voxels selected as informative for classification.

We predicted that activation vectors associated with conscious perception (i.e. visible stimuli) would exhibit less trial-to-trial variability in their angle than those associated with non-conscious perception (reflecting greater reproducibility), without necessarily any difference in the norm (i.e. intensity). To assess the reproducibility of representations, we measured the variability in the angle between pairs of vectors (both from the same run and same stimulus category), as well as the norm of each vector, separately for visible and invisible stimuli (*13, 21*). We repeated this in both the "visible" and the "invisible" selections (*22*). This resulted in four sets of data: responses to visible and invisible stimuli in the "visible" selection, and responses to visible and invisible stimuli in the "invisible" selection. To avoid confounds likely to arise from comparing

properties of vectors in different subsets of voxels (and hence different regions of cortex), we restricted our comparisons to vectors within the same selection (*23*). We used the mean within-category within-run angular deviation as an index of reproducibility.

Figure 2B shows that, within the "invisible" selection, the variability of the vector angle (dVA) is significantly less for visible than for invisible stimuli ($p < 0.01$, paired-samples two-sided signed rank test). There was no difference in dVA between visible and invisible stimuli in the "visible" selection (Fig. 2A), suggesting that the variability is found primarily in the subset of voxels that carry non-conscious information, and that this subset is distinct from that within which conscious information is found (for this particular combination of stimuli and task). This is consistent with the failure of the classifier to generalize across the two levels of visibility. When dVA for the "invisible" selection was compared with the baseline level 4 seconds prior (i.e. at the time of stimulus onset), there was a significant interaction ($p = 0.021$, two-sided signed rank test on the deviation from baseline): dVA is below baseline in response to visible stimuli and is higher than baseline in response to invisible stimuli (Fig. 2B). There was no difference in the mean or variance of the vector norm for visible versus invisible stimuli, either in the "visible" or "invisible" selection (Fig. 2C & D; means: $p > 0.35$, paired-samples two-sided signed rank test; variances: $p > 0.7$, Levene's test). Thus a difference in signal to noise ratio is not sufficient to explain the effect.

Since measurable category-specific information had been identified separately for both visible and invisible stimuli, we examined where in the brain the information tended to coalesce in each case (Fig. 3). For any given subject, reliably informative voxels could be found throughout the

temporal lobes (Fig. 3A). Averaging across subjects (*24*) revealed two clusters in the right ventral temporal cortex, one for visible and the other for invisible stimuli, with minimal spatial overlap, consistent with the failure of the classifier trained on one type of stimulus to generalize to the other (Fig. 3B, C). The anterior-posterior relationship of the two clusters ("visible" and "invisible" selections, respectively) coincides with previous observations (*25*).

Conscious and non-conscious neural activation patterns coexist within the cerebral cortex, side by side at the same time, but presumably they differ in several ways. Proposed differences include duration, intensity, and coherence. Here we show that they also differ in their relative reproducibility across presentations of similar stimuli. Why might reproducibility distinguish conscious from non-conscious representations? One possibility is that conscious information is represented in a more discrete form (*26*), making it more durable and robust, but also more stereotypical (and therefore more reproducible). Another possibility is that conscious information manifests itself in relatively stable neural firing patterns, corresponding to the "settled" states of recurrent network interactions (*27*). There are a number of plausible theories regarding the neural correlates of consciousness, but relatively little data concerning the *nature* of conscious versus non-conscious encoding. Further work is required to understand the difference(s) in the way perceptual information is encoded in the brain depending on whether or not that information is present in subjective experience. Such work is likely to have profound importance in a variety of arenas, including the assessment of consciousness under presumed anaesthesia or coma and the investigation of brain function in conditions such as schizophrenia, autism, and dissociation disorders.

References and Notes

Supporting Online Material

www.sciencemag.org

Materials and Methods

Supplementary figure S1

Tables

TRAIN \ TEST	VISIBLE	INVISIBLE
VISIBLE	63 +/- 3.5 t=3.8, p<0.002*	48 +/- 2.3 t=-0.78, p=0.77
INVISIBLE	52 +/- 3.0 t=0.69, p=0.25	58 +/- 3.1 t=2.5, p<0.02*

Table 1: Performance of a Gaussian naïve-Bayes classifier

The objective of the classifier was to discriminate the category of the stimulus based on the pattern of beta weights (a GLM was applied separately to each run of functional data, see SOM). A voxel-wise ANOVA and nested cross-validation (*18*) were used for dimensionality reduction on each round of training/testing. For within-condition classification (i.e. visible-visible, invisible-invisible) a leave-one-run-out cross-validation was performed. For between-condition classification we trained on all the data from one condition and tested on the other, and vice-versa. All t-tests are one-tailed with df = 11.

Figures legends

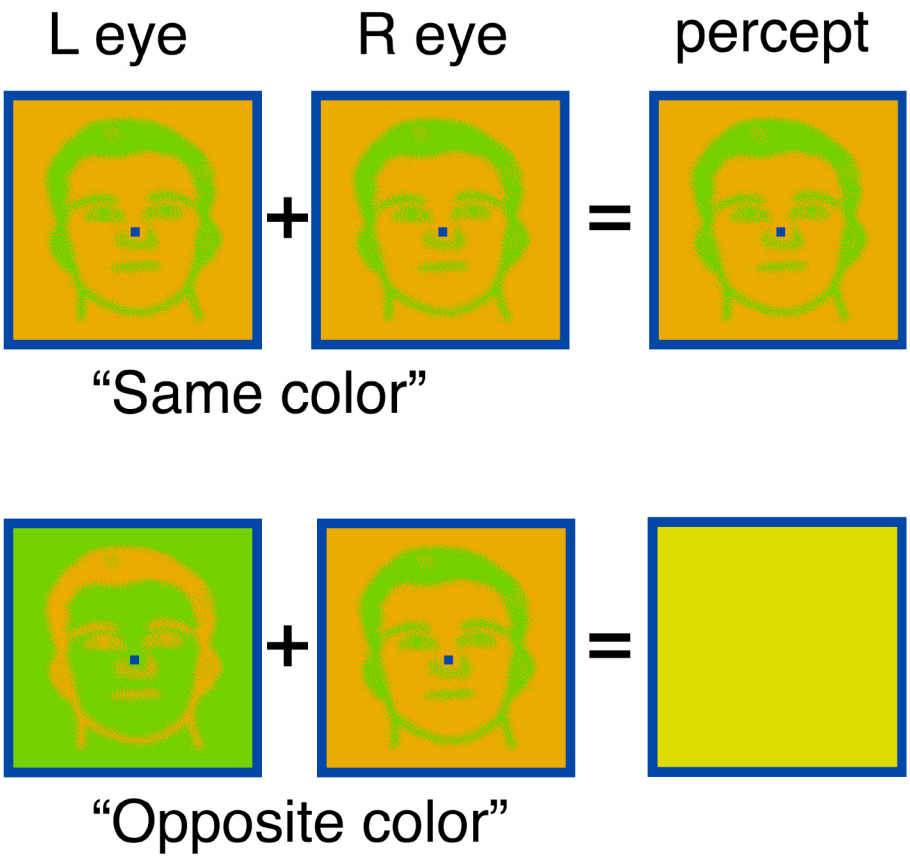

Fig. 1: Dichoptic-color masking

This method of manipulating awareness, originally devised by (*7*), relies on the phenomenon of dichoptic color fusion. The "same color" mode corresponds to the "visible" condition and the "opposite color" mode corresponds to the "invisible" condition. In order to achieve disappearance of the image in the "opposite color" mode, the two colors must be approximately isoluminant and the object boundaries slightly blurred. Before the experiment, subjects were trained to maintain steady fixation, and were cued to do so during each trial with the appearance of the fixation point (500ms before stimulus onset). Stimuli were presented stereoscopically in the fMRI scanner using a cardboard divider and prism lenses (*28*).

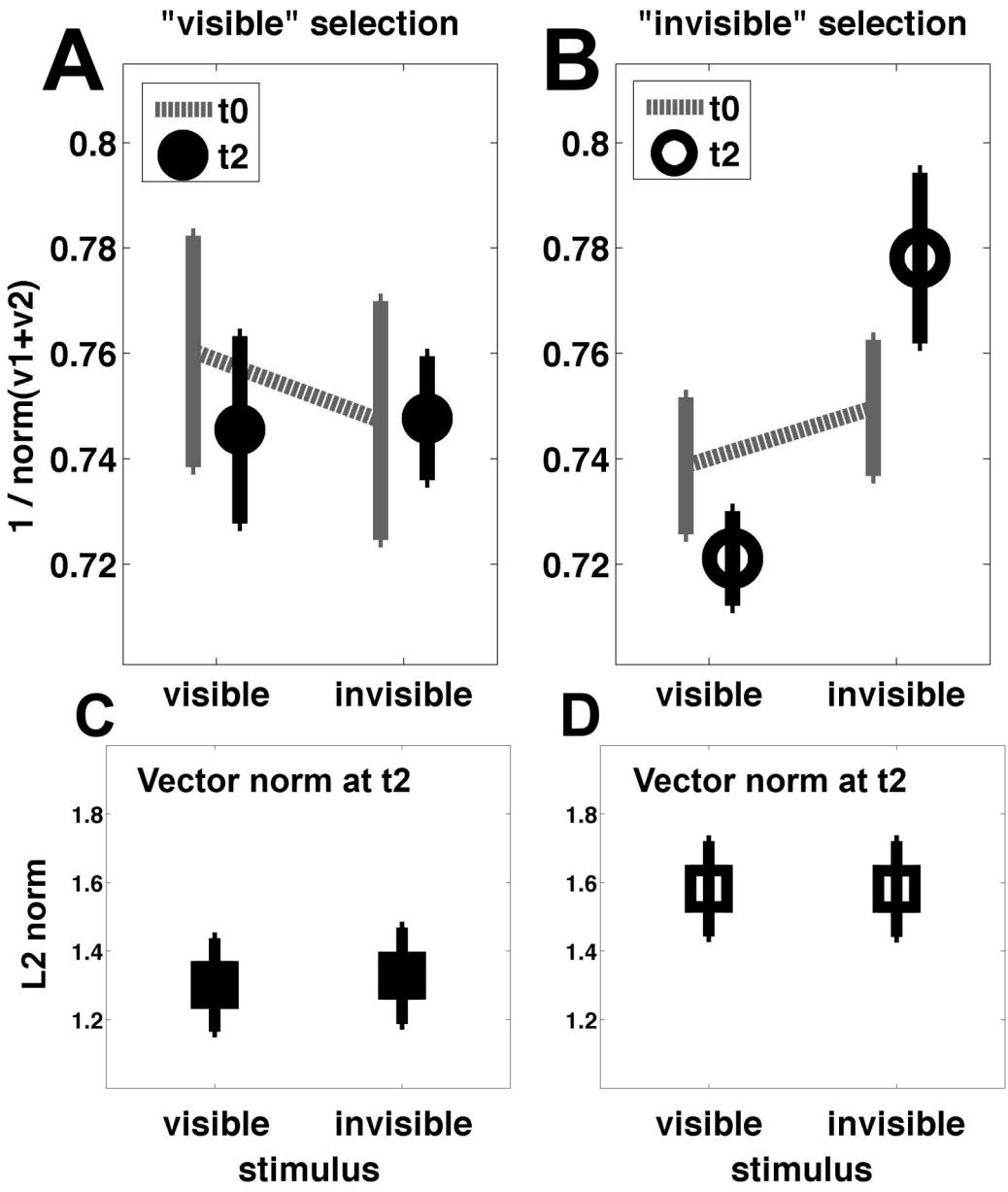

Fig. 2: Variability in the angle of activation vectors in the "visible" and "invisible" selections (A, B), and mean vector norm (C, D).

In both A and B, ***t0*** corresponds to the TR (repetition time = 2 sec) on which the stimulus was presented, before the haemodynamic response had begun to rise. ***t2*** corresponds to 2 TR's (4 seconds) after the stimulus was presented, at the (approximate) peak of the haemodynamic response. N = 12 subjects. This analysis was performed using a leave-one-run-out procedure: voxel selection was performed on data from n-1 runs, and the norm and angular deviation were computed on data from the run that had been left out (see SOM). Comparisons between the two selections (A versus B or C versus D) are not valid (*23*).

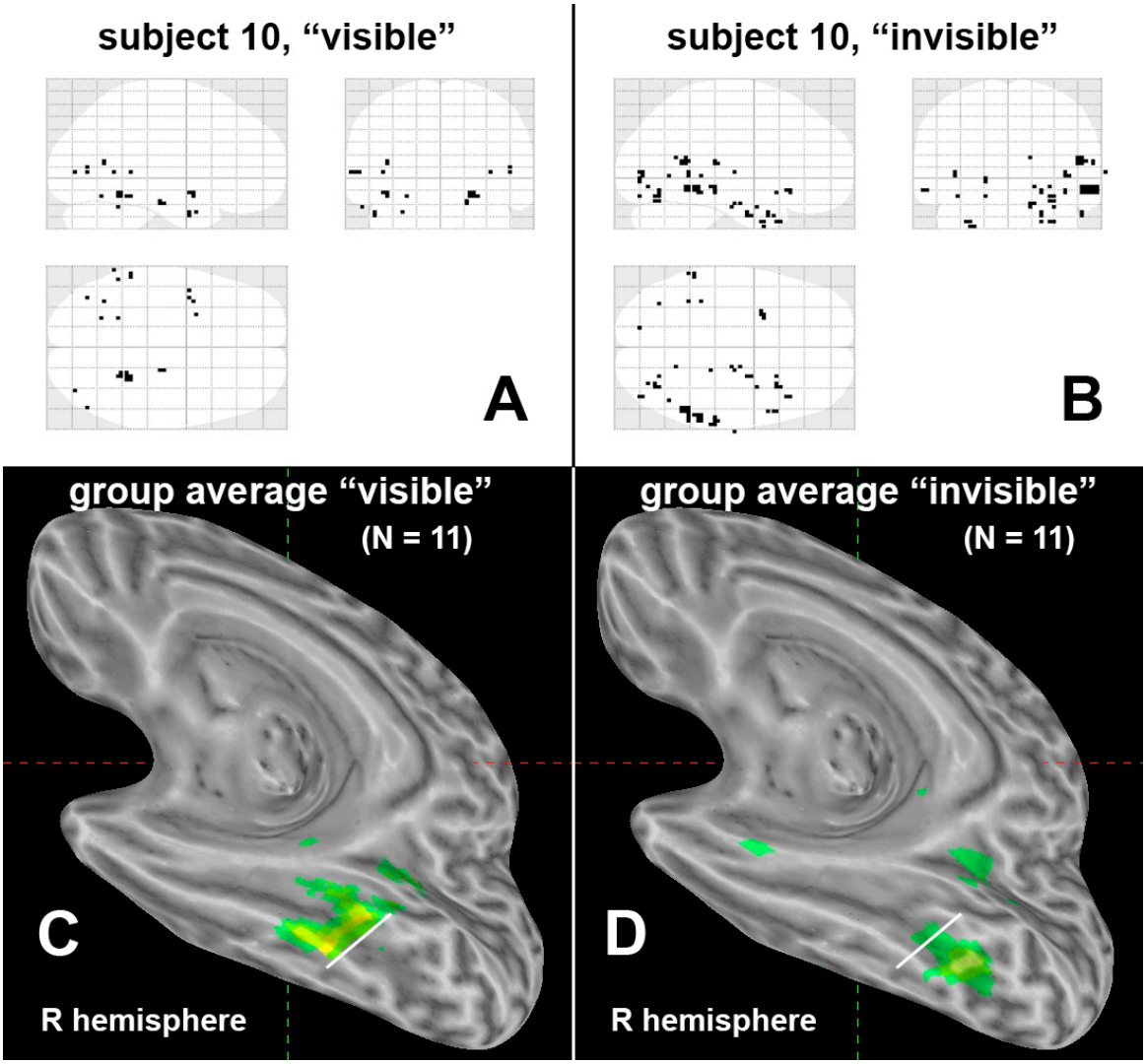

Fig. 3: Spatial distribution of informative voxels

A and B show voxels that were selected as informative for classification (face versus house) on 6 or more (out of 12) runs, for a subject with comparable classification accuracy (72% correct) for visible and invisible stimuli. C and D show the mean across subjects (*24*) projected onto the AFNI TT_N27 template brain (right hemisphere) at a statistical threshold of $p < 0.05$ (corrected). The oblique white line serves as a visual landmark. The cluster in C encompasses a portion of the fusiform and parahippocampal gyri, in the area of the fusiform face area (FFA) and parahippocampal place area (PPA). The cluster in D lies along the posterior fusiform gyrus.

References and Notes

1. S. Dehaene *et al.*, *Nature* **395**, 597 (1998).
2. P. M. Merikle, D. Smilek, J. D. Eastwood, *Cognition* **79**, 115 (2001).
3. S. Kouider, S. Dehaene, *Philosophical Transactions of the Royal Society B-Biological Sciences* **362**, 857 (2007).
4. Y. Jiang, S. He, *Curr Biol* **16**, 2023 (2006).
5. G. Rees *et al.*, *Brain* **123**, 1624 (2000).
6. P. Vuilleumier *et al.*, *Proc Natl Acad Sci USA* **98**, 3495 (2001).
7. K. Moutoussis, S. Zeki, *Proc Natl Acad Sci USA* **99**, 9527 (2002).
8. L. Pessoa, S. Kastner, L. G. Ungerleider, *J Neurosci* **23**, 3990 (2003).
9. B. Bahrami, N. Lavie, G. Rees, *Current Biology* **17**, 509 (2007).
10. V. Wyart, C. Tallon-Baudry, *J. Neurosci.* **28**, 2667 (2008).
11. A. Schurger, A. Cowey, J. D. Cohen, A. Treisman, C. Tallon-Baudry, *Neuropsychologia* **46**, 2189 (2008).
12. M. Kinsbourne, in *Scientific approaches to consciousness,* J. D. Cohen, J. W. Schooler, Eds. (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1997), pp. 335-355.
13. Materials and methods are available as supporting material on Science Online.
14. Post-decision wagering has been proposed as an independent measure of awareness, under the assumption that if the subject is aware of the relevant sensory information then there will be a correspondence between high wagers and correct responses. If no such correspondence is found, then under this assumption we deduce that the subject was not aware of the relevant sensory information (in this case, information sufficient to discern the category of the stimulus). See {ref. Persaud}.
15. N. Persaud, P. McLeod, A. Cowey, **10**, 257 (2007).
16. A. Schurger, S. Sher, *TICS* **12**, 209 (2008).
17. J. D. Haynes, *Trends Cogn Sci* **13**, 194 (2009).
18. F. Pereira, T. Mitchell, M. Botvinick, *Neuroimage* **45**, S199 (2009).
19. D. L. Sheinberg, N. K. Logothetis, *Proc Natl Acad Sci USA* **94**, 3408 (1997).
20. P. Sterzer, J. D. Haynes, G. Rees, *J Vis* **8**, 10 1 (2008).
21. While voxel selection was based on coefficients derived statistically from each functional run (see SOM for details), the activation patterns among these voxels were taken trial by trial from the minimally-processed fMRI signal data (at t0 + 2TR, where t0 is time of stimulus onset and 1 TR = 2 seconds). This was done in a leave-one-out fashion: the selection was chosen based on data from n-1 runs, and then the activity vectors from the left-out run (2 @ visible / invisible x face / house per run) were projected into that space (see SOM for details).
22. The "visible selection" comprises the voxels that were maximally informative as to the category of visible stimuli. Likewise, the "invisible selection" comprises the voxels that were maximally informative as to the category of invisible stimuli.
23. Since the "visible selection" and the "invisible selection" occupy separate and largely non-overlapping regions of cortex, then comparisons between their functional properties are confounded with differences between the haemodynamic and magnetic-field properties of the regions that they inhabit.

24. To produce spatial maps of reliably informative voxels, each voxel was coded with either a ‘1’, if selected on a majority of runs, or a ‘0’ otherwise (Fig. 3, A and B). In order to uncover regional tendencies in the average across subjects, maps for each subject were blurred by ~ 10mm and then discretized again (ceiling) . The probability distribution of the average map under the null hypothesis was estimated using a permutation test (number of voxels held constant for each subject / selection, but locations randomized) and used to set a statistical threshold.
25. M. Bar *et al.*, *Neuron* **29**, 529 (2001).
26. J. Sackur, S. Dehaene, *Cognition* **111**, 187 (2009).
27. D. Balduzzi, G. Tononi, *PLoS Comput Biol* **4**, e1000091 (2008).
28. A. Schurger, *J Neurosci Methods* **177**, 199 (2009).
29. A.S. was supported by a grant from the Mind Science Foundation and by a Ruth L. Kirschstein National Research Service Award from the NIMH (MH075342). Special thanks to Shlomi Sher for helpful discussions, to Stanislas Dehaene and two anonymous reviewers for comments, to Minsoo Kim for help with behavioral testing, and to Leigh Nystrom for advice and assistance with data analysis.