### Reporting Checklist for Nature Neuroscience

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. For more information, please read Reporting Life Sciences Research.

Please note that in the event of publication, it is mandatory that authors include all relevant methodological and statistical information in the manuscript.

#### Statistics reporting, by figure

- Please specify the following information for each panel reporting quantitative data, and where each item is reported (section, e.g. Results, & paragraph number).
- Each figure legend should ideally contain an exact sample size (n) for each experimental group/condition, where n is an exact number and not a range, a clear definition of how n is defined (for example x cells from x slices from x animals from x litters, collected over x days), a description of the statistical test used, the results of the tests, any descriptive statistics and clearly defined error bars if applicable.
- For any experiments using custom statistics, please indicate the test used and stats obtained for each experiment.
- Each figure legend should include a statement of how many times the experiment shown was replicated in the lab; the details of sample collection should be sufficiently clear so that the replicability of the experiment is obvious to the reader.
- For experiments reported in the text but not in the figures, please use the paragraph number instead of the figure number.

Note: Mean and standard deviation are not appropriate on small samples, and plotting independent data points is usually more informative. When technical replicates are reported, error and significance measures reflect the experimental variability and not the variability of the biological process; it is misleading not to state this clearly.

<table>
<thead>
<tr>
<th>Fig. number</th>
<th>Test used</th>
<th>Section &amp; paragraph #</th>
<th>Exact value</th>
<th>Defined?</th>
<th>REported?</th>
<th>Descriptive stats (average, variance)</th>
<th>P value</th>
<th>Degrees of freedom &amp; f/t/z/r/etc value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a</td>
<td>one-way ANOVA</td>
<td>Methods para 8</td>
<td>9, 9, 10, 15 mice from at least 3 litters/group</td>
<td>error bars are mean +/- SEM</td>
<td>Fig. legend</td>
<td>p = 0.044</td>
<td>F(3, 36) = 2.97</td>
<td>Fig. legend</td>
</tr>
<tr>
<td>3</td>
<td>unpaired t-test Results para 6</td>
<td>Results para 6</td>
<td>15 slices from 10 mice</td>
<td>error bars are mean +/- SEM</td>
<td>Results para 6</td>
<td>p = 0.0006</td>
<td>t(28) = 2.808</td>
<td>Results para 6</td>
</tr>
<tr>
<td>TEST USED</td>
<td>n</td>
<td>DESCRIPTIVE STATS (AVERAGE, VARIANCE)</td>
<td>P VALUE</td>
<td>DEGREES OF FREEDOM &amp; F/T/Z/R/ETC VALUE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>---</td>
<td>--------------------------------------</td>
<td>---------</td>
<td>---------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FIGURE NUMBER</strong></td>
<td><strong>WHICH TEST?</strong></td>
<td><strong>SECTION &amp; PARAGRAPH #</strong></td>
<td><strong>EXACT VALUE</strong></td>
<td><strong>DEFINED?</strong></td>
<td><strong>REPORTED?</strong></td>
<td><strong>SECTION &amp; PARAGRAPH #</strong></td>
<td><strong>EXACT VALUE</strong></td>
<td><strong>SECTION &amp; PARAGRAPH #</strong></td>
</tr>
<tr>
<td>3a</td>
<td>bootstrap</td>
<td>All statistical methods are described on Page 24, in the Methods subsection entitled &quot;Statistical Methods&quot;</td>
<td>100</td>
<td>For both IT and V4 populations, performances of most informative site for each of several tasks, as indicated in the figure labeling.</td>
<td>Error bars are 68% confidence intervals (CI) centered at the median for each bar. CI variance is due to sampling images used to determine maximum-information site. For IT, CI also includes variation due to random sampling of 126 sites to match V4 population size.</td>
<td>Methods &quot;Neural Performance Assessment&quot; Para. 1 and Methods sec. &quot;Statistical Methods&quot;</td>
<td>All tasks and both areas (IT, V4) are statistically different from 0 below the p = 0.005 confidence level.</td>
<td>Methods &quot;Neural Performance Assessment&quot; Para. 1 and Methods sec. &quot;Statistical Methods&quot; and Supplementary Statistical Table (Supplementary Item 15) for all p-values.</td>
</tr>
<tr>
<td>3a</td>
<td>paired t-test</td>
<td>p 24</td>
<td>Differences in performances between IT and V4 most informative site. These were paired using the sets of images used to determine the most informative sites.</td>
<td>Error bars are mean +/- SEM</td>
<td>The following is a listing of specific values for tasks for which the IT-V4 difference is not significant at the 0.005 level: Y-Axis position p = 0.012 Y-Axis Size p = 0.032 2-D Retinal Area p = 0.075 Major Axis Angle p = 0.045 X-axis rotation p = 0.366 Y-Axis rotation p = 0.135 Z-Axis rotation p = 0.241</td>
<td>Methods &quot;Neural Performance Assessment&quot; Para. 1 and Methods sec. &quot;Statistical Methods&quot;</td>
<td>Y-Axis position t = 2.63 Y-Axis Size t = 2.215 Major Axis Angle t = 2.069</td>
<td>Methods &quot;Neural Performance Assessment&quot; Para. 1 and Methods sec. &quot;Statistical Methods&quot;</td>
</tr>
<tr>
<td>+ 3b</td>
<td>bootstrap</td>
<td>p 24</td>
<td>100</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------------</td>
<td>------</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>For IT, V4, V1-like model and pixel control, these are median over bootstraps of population decoding values for each of 16 listed tasks.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Methods**

"Neural Performance Assessment" Para. 1 and Methods sec. "Statistical Methods"

**Error bars**

Error bars are 68% confidence intervals (CI) centered at the median for each bar. CI variance is due to sampling images used to build linear decoders. For IT, V1-like and pixel populations, CI also includes variation due to random sampling of 126 sites to match V4 population size. All tasks and both areas (IT, V4) are statistically different from 0 below the p = 0.005 confidence level. The following is a listing of specific p-values for tasks where the V4-V1like separation was not significant at the 0.005 level:

- X-axis position: p = 0.0625
- Y-Axis Size: p = 0.0243
- Y-Axis Rotation: p = 0.0495
- X-axis Rotation: p = 0.396

The following is a listing of specific t-statistics for tasks for which the V4-V1 performance difference is significant at the 0.05 level but not significant at the 0.005 level:

- Major Axis Angle: t = 2.382
- Y-Axis Size: t = 2.783
- Y-Axis Rotation: t = 2.173

All other t-statistics were either larger than 3.0 (when IT-V4 or V4-V1 differences were significant at the p=0.005 level or smaller than 2, when differences were not significant at the p=0.05 level.)

**N/A for bootstrap test.**

"Neural Performance Assessment" Para. 1 and Methods sec. "Statistical Methods"
<p>| # | 7a | bootstrap | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; | 100 | Same as bootstrap for figure 2b. | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; | Same as bootstrap for figure 2b. | For all tasks, V1 and IT features are statistically different from 0 with a p-value of less than 0.005. The same is true for the pixel representation for the position tasks. For pixels, on orientation, difference from 0 is not significant at the 0.05 level, with a p-value of 0.075. For the V4 population: X-position p-value = 0.046 Y-position p-value = 0.038 Orientation p-value = 0.413. | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; and Supplementary Statistical Table (Supplementary Item 15) for all p-values. | N/A | N/A |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # | 7a | paired t-test | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; | 100 | Same as t-test for figure 2b. | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; | The following is a list of specific p-values for tasks where the IT-V4 difference was not significant at the 0.05 level: X-Position p = 0.318 Y-Position p = 0.092 | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; and Supplementary Statistical Table (Supplementary Item 15) for all p-values. | All other t-statistics were either larger than 3.0 (when IT-V4 or V4-V1 differences were significant at the p=0.005 level or smaller than 2, when differences were not significant at the p=0.05 level. | Methods &quot;Performance in Simple Stimuli&quot; and Methods sec. &quot;Statistical Methods&quot; and Supplementary Statistical Table (Supplementary Item 15) for all p-values. |</p>
<table>
<thead>
<tr>
<th>Method</th>
<th>4b (left and right pane l)</th>
<th>1-way ANOVA</th>
<th>100</th>
<th>Same as above.</th>
<th>N/A</th>
<th>N/A</th>
<th>p &lt; 1e-5</th>
<th>N/A</th>
<th>Methods “Human Psychophysical Experiments” paras. 6 and 7 and Methods sec. “Statistical Methods”</th>
<th>Supplem entary Statistical Table (Supplementary Item 15) for all p-values.</th>
</tr>
</thead>
<tbody>
<tr>
<td>bootstrap</td>
<td>Methods “Human Psychophysical Experiments” paras. 6 and 7 and Methods sec. “Statistical Methods”</td>
<td>100</td>
<td>Four populations were subject to the bootstrap: for each of the IT, V4, V1like and pixel populations, we computed r-values of performances for linear decoders built on the populations, with the human performances, across sets of tasks. We also computed bootstraps for the human population.</td>
<td>Error bars are 68% confidence intervals (CI) centered at the median for each population. CI variance is due to sampling images used to build linear decoders and tasks used to compute the r-value. For IT, V1like and pixel populations, CI also includes variation due to random sampling of 126 sites to match V4 population size. CI for the human population (horizontal gray bar in 5b panels) was computed by bootstrapping over population splits and tasks.</td>
<td>Methods “Human Psychophysical Experiments” paras. 6 and 7 and Methods sec. “Statistical Methods”</td>
<td>N/A</td>
<td>p &lt; 1e-5</td>
<td>Methods “Human Psychophysical Experiments” paras. 6 and 7 and Methods sec. “Statistical Methods” and Supplementary Statistical Table (Supplementary Item 15) for all p-values.</td>
<td>N/A for bootstrap test.</td>
<td></td>
</tr>
<tr>
<td>Page</td>
<td>Bootstrap</td>
<td>Method</td>
<td>Number</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-----------</td>
<td>--------</td>
<td>--------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6c</td>
<td>bootstrap</td>
<td>Methods &quot;Evaluation on the Testing Set&quot;</td>
<td>100</td>
<td>For each of 15 non-categorization tasks, bars show the median over bootstrap splits of correlation of performance during categorization training of the categorization task performance with the indicated non-categorization task.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6e</td>
<td>bootstrap</td>
<td>Methods &quot;Evaluation on the Testing Set&quot;</td>
<td>100</td>
<td>For each layer of the model, bar height represents median over bootstraps of model-neural performance correlation over tasks.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7b</td>
<td>bootstrap</td>
<td>Methods &quot;Evaluation on the Testing Set&quot;</td>
<td>500</td>
<td>For each of three model layers, bar height represents median over bootstraps of model-layer performance on indicated task.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Error bars are 68% confidence intervals centered at the median for each layer. CI variance is due to variation in images used to build linear decoders as well as subsampling in units from model layer. All correlations are different from 0 at a p-value of less than 1e-10.

N/A. N/A

Methods "Evaluation on the Testing Set" Para. 1 and Methods sec. "Statistical Methods" and Supplementary Statistical Table (Supplementary Item 15) for all p-values.

Methods "Evaluation on the Testing Set" | Para. 1 and Methods sec. "Statistical Methods" | N/A | N/A

Methods "Evaluation on the Testing Set" | Para. 1 and Methods sec. "Statistical Methods" | All (non-pixel) bars are significantly different from 0 at <0.001 confidence level.

N/A. N/A

Methods "Evaluation on the Testing Set" Para. 2 and Methods sec. "Statistical Methods" and Supplementary Statistical Table (Supplementary Item 15) for all p-values.
**7b** paired t-test

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods"

Differences in performances between Layer1-Layer3 and Layer3-Layer6 model populations. These were paired using the sets of images used to build the linear decoder.

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods"

error bars are mean +/- standard deviation of the mean

Differences are significant at below 0.005 level unless specified. (*) means different at 0.05 level but not 0.01. n.s. means not significant at 0.05 level.

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods" and Supplementary Table Item 15 for all p-values.

---

**7c -- basic categorization**

paired t-tests

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods"

Differences between performances in IT-V4 and model layer6-layer3 comparison at left-most end of rotational variation spectrum.

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods"

error bars are mean +/- standard deviation of the mean

Differences between IT and V4 is not significant at 0.05 level. Difference between layer 3 and layer 6 is not significant at 0.05 level.

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods" and Supplementary Table Item 15 for all p-values.

---

**7c -- subordinate identification**

paired t-tests

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods"

Differences between performances in IT-V4 and model layer6-layer3 comparison at left-most end of rotational variation spectrum.

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods"

error bars are mean +/- standard deviation of the mean

V4 is greater than IT with significance a < 0.001 level. Model layer 3 is greater than model layer 6 with significance at < 0.001 level.

Methods
"Evaluation on the Testing Set" Para 2 and Methods sec. "Statistical Methods" and Supplementary Table Item 15 for all p-values.
Methods
"Evaluation on the Testing Set" Para. 2 and Methods sec. "Statistical Methods"

Differences between performances in IT-V4 and model layer6-layer3 comparison at left-most end of rotational variation spectrum.

error bars are mean +/- standard deviation of the mean

IT is greater than V4 with confidence at < 0.001 level, and same with model Layer 6 performance being greater than Layer 3.

V4 performance is greater than IT performance with confidence at < 0.005 level.

Model Layer 6 and layer 3 performance are not significantly different at 0.05 level.

Supplementary Table (Supplementary Item 15) for all p-values.

<table>
<thead>
<tr>
<th>7c -- y-axis position</th>
<th>paired t-tests</th>
<th>Methods &quot;Evaluation on the Testing Set&quot; Para 2 and Methods sec. &quot;Statistical Methods&quot;</th>
<th>Differences between performances in IT-V4 and model layer6-layer3 comparison at left-most end of rotational variation spectrum.</th>
<th>error bars are mean +/- standard deviation of the mean</th>
<th>IT is greater than V4 with confidence at &lt; 0.001 level, and same with model Layer 6 performance being greater than Layer 3.</th>
<th>V4 performance is greater than IT performance with confidence at &lt; 0.005 level. Model Layer 6 and layer 3 performance are not significantly different at 0.05 level.</th>
<th>Methods &quot;Evaluation on the Testing Set&quot; Para. 2 and Methods sec. &quot;Statistical Methods&quot; and Supplementary Table (Supplementary Item 15) for all p-values.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7c -- 3d object scale</td>
<td>paired t-tests</td>
<td>Methods &quot;Evaluation on the Testing Set&quot; Para 2 and Methods sec. &quot;Statistical Methods&quot;</td>
<td>Differences between performances in IT-V4 and model layer6-layer3 comparison at left-most end of rotational variation spectrum.</td>
<td>Methods &quot;Evaluation on the Testing Set&quot; Para 2 and Methods sec. &quot;Statistical Methods&quot;</td>
<td>error bars are mean +/- standard deviation of the mean</td>
<td>V4 performance is greater than IT performance with confidence at &lt; 0.005 level. Model Layer 6 and layer 3 performance are not significantly different at 0.05 level.</td>
<td>Methods &quot;Evaluation on the Testing Set&quot; Para. 2 and Methods sec. &quot;Statistical Methods&quot; and Supplementary Table (Supplementary Item 15) for all p-values.</td>
</tr>
</tbody>
</table>

Supplementary Figure 11

Same as main text Figure 6a

Same as main text Figure 5c, namely correlation between performance on categorization and indicated task in the test neural data set.

Bar height is same as in main text Figure 5c, namely correlation between performance on categorization and indicated task in the test neural data set.

500

Correlations for Layers 1, 3, and 4 are different from 0 at 0.01 level.

Correlation for layer 6 is not different from 0 at 0.05 level. (p = 0.14, t = -1.24)

Same as main text Figure 5c

Same as main text Figure 6a

Correlation for Layers 1, 3, and 4 are different from 0 at 0.01 level.

Correlation for layer 6 is not different from 0 at 0.05 level. (p = 0.14, t = -1.24)

Same as main text Figure 6a

Bootstrap

Same as main text Figure 5c

500

Error bars 68% confidence intervals centered at the median for each task. CI variance is due to variation in images used to build linear decode.

Correlations for Layers 1, 3, and 4 are different from 0 at 0.01 level.

Correlation for layer 6 is not different from 0 at 0.05 level. (p = 0.14, t = -1.24)

Same as main text Figure 5c

Correlation for Layers 1, 3, and 4 are different from 0 at 0.01 level.

Correlation for layer 6 is not different from 0 at 0.05 level. (p = 0.14, t = -1.24)

Same as main text Figure 6a

Bootstrap

Same as main text Figure 5c

500

Error bars 68% confidence intervals centered at the median for each task. CI variance is due to variation in images used to build linear decode.

Correlations for Layers 1, 3, and 4 are different from 0 at 0.01 level.

Correlation for layer 6 is not different from 0 at 0.05 level. (p = 0.14, t = -1.24)
Representative figures

1. Are any representative images shown (including Western blots and immunohistochemistry/staining) in the paper?

   If so, what figure(s)?

   Fig 6, panels A and C, show representative examples for several tasks of the dependence of the performance of the model on training-set categorization performance (B) and model layer (A). These tasks were chosen to illustrate the point that the results are similar for quite different tasks (position vs scale vs pose). Equivalent images for all measured tasks are shown in supplementary figures 11 and 12.

   Fig. 7 c's panels are representative in that they show 4 examples out of 16 tasks. These tasks were chosen to illustrate the points that (a) at low amounts of rotational variation, V4 and IT performance gaps close and (b) for some tasks, the gap reverse (e.g. IT has lower performance at V4). Equivalent images for all measured tasks are shown in supplementary figure 7.

   Supplementary Figures 2 panes A, B, and C are representative images. They show the average response properties of specific chosen neural sites from our data. In each panel, the sites chosen were those that had the maximal amount of information for categorization, position, and object size tasks, respectively. These sites were selected by by the same procedure used in the quantification main test Figure 3A, namely: out of the set of all 266 IT units that we collected, for each task, we computed the performance metric on a subset of the images and selected the best sites for the task; we then show average response values on a held-out set of validation images not used to select the sites.

   Supplementary Fig 5 panel A contains representative images of distributions of decoder weights for several tasks. We chose these images by picking those with the highest, lowest, and closest-to-median levels of sparseness.

2. For each representative image, is there a clear statement of how many times this experiment was successfully repeated and a discussion of any limitations in repeatability?

   If so, where is this reported (section, paragraph #)?

   For the purposes of selecting all representative images in main text figure 7A,B, the selection procedure was performed only once, with the standard parameters used elsewhere in the work. We don’t report this as such anywhere in the task.
Statistics and general methods

1. Is there a justification of the sample size?
   If so, how was it justified?
   Where (section, paragraph #)?
   Even if no sample size calculation was performed, authors should report why the sample size is adequate to measure their effect size.

   There are several sample sizes that arise in this paper:

   a. Number of animals used: In our electrophysiology experiments, we collect data from two rhesus macaque monkeys. This is a common sample size that is frequently used in the field of monkey physiology.

   b. Number of neuronal sites sampled: Between the two animals, we collected data from a total of 392 visually-driven neural sites, a sample size that is comparatively large relative to other electrophysiology studies using similar techniques. This sample size is justified in each result by always ensuring that for results in which neuronal sample size would be relevant (Figs. 3 - 5), error bars were computed over variation that included subsampling of neuronal sites. Moreover, we estimate the robustness of these results by using randomly subsampling of sites to show that our main results hold by a significant margin for *every* sample size between 1 and our total (see Fig. 4a and Supp. Fig. 4).

   c. Number of repetitions per image for each neuronal site. We recorded between 25 and 50 repetitions of each image for each neuronal site, as discussed in the section on Array Electrophsiology in the methods, page 16. Prior to the experiment, we chose this number by experience from previous studies (both in our lab and external) that suggested that this number of repetitions would be more than sufficient to obtain a stable estimate of the mean response of any given IT or V4 neuron to each visual image. Ex post facto, we found that for the typical V4 or IT site, the self-consistency of the responses of that site was quite high (approximately 0.8). As a result, we likely could have obtained highly reliable estimates of mean firing rates with fewer than 20 repetitions, as confirmed by further analysis that we did plotting the reliability of the estimate as a function of the number of repetitions (subsampled from our 25-50 reps).

   d. Number of human test subjects: In our human psychophysical tests, for each task, we collected data from 80 subjects, which is about the norm (or slightly above it) for comparable psychophysical studies. We also considered the levels of subject-to-subject consistency for each task, described in methods section "human psychophysical experiments", in a separate paragraph for each task. Given that these consistency values are higher than 0.85 in nearly all cases, we did perform a separate sample size calculation.

   e. Number of image stimuli collected: Data from both humans and macaques were collected on 5,760 distinct stimuli. This number of stimuli was justified by computing error bars over variation that included subsampling images (Fig. 3d, 4, 5, and 7). Moreover, we estimate the robustness of these results by using randomly subsampling images to show that our main results hold by a significant margin for a large range of images sample sizes (see Supp. Figs. 2).

   f. Number of tasks: We justify the number of tasks chosen by ensuring that in results that depended on this choice (Fig. 4), error bars were computed over variation that included subsampling in the set of tasks.
2. Are statistical tests justified as appropriate for every figure?

Where (section, paragraph #)?


a. If there is a section summarizing the statistical methods in the methods, is the statistical test for each experiment clearly defined?

Yes.

b. Do the data meet the assumptions of the specific statistical test you chose (e.g. normality for a parametric test)?

The t-tests used depended on independent observations for each population being compared, conditions that were true by construction of our subsampling processes.

The ANOVA used in Fig. 4 did required normality, a condition that was approximately true by construction of our subsampling processes.

c. Is there any estimate of variance within each group of data? Is the variance similar between groups that are being statistically compared?

- Yes. Using bootstrapping methods, we computed a standard 68%-confidence interval was given as the variability estimation for each element in fig 3, Fig 4, Fig 6, and Fig 7.

- Usually variance was similar between groups (as one can see by visual inspection of figures), but we used methods that allowed for different variances (e.g. Welch F-Test in every comparison).

- Described in "statistical methods" subsection of the Methods, page 24.

d. Are tests specified as one- or two-sided?

When t-tests were used (Fig 3d, Fig 4, Fig 6, Fig 7), they were two-sided.

e. Are there adjustments for multiple comparisons?

In figure 4c, we used a standard 1-way ANOVA to first determine whether group averages between each group where different; they were. We then applied a t-test to each group to determine the significance of differences between each group and the human population average.

In Fig. 7c, the statistical significance of the fact that V4 can have more information than IT for some tasks in low-variation conditions, should be corrected using the Bonferroni multiple comparison criteria, where the number of tests is the number of tasks. After all, our claim is not merely that this specific task sees a V4/IT reversal at low variation levels, but rather that there IS such a task at all. In theory we could have cherry-picked this task (though see Supp. Fig. 8 for some other examples.) However, given that we tested 16 tasks, the magnitude of the difference here, and the error bars, this claim still survives Bonferroni correction with M = 16.
3. Are criteria for excluding data points reported?  
   Was this criterion established prior to data collection?  
   Where is this described (section, paragraph #)?

   a. In the analysis of our electrophysiology recordings, we removed sites from our original neural sample sites that were not visually responsive. The reason that we were required to do this is that the chronically-implanted multi-electrode array technique that we used does not allow us to move the electrodes once the device has been implanted. In many cases, individual electrodes were not situated close to visually responsive neurons, and so did not give reliably modulated responses across stimuli. As described in the methods section "Array Electrophysiology" (page 29), sites were selected as being visually driven using a standard imageset distinct from the one on which we performed the remainder of our analysis. From our nine 96-electrode arrays, a total of 392 sites were deemed visually driven.

   b. As discussed in the Methods subsection on "Human psychophysical experiments", we occasionally excluded from our psychophysical data subjects whose response patterns indicated that they were cheating on the task by giving identical responses across most or all trials. This criteria for exclusion was established in previous work in our research group, prior to this study being commenced. For each of our tasks, each having a total of 80 subjects, we never excluded more than 2 subjects -- though for most tasks, no subjects ended up being excluded.

4. Define the method of randomization used to assign subjects (or samples) to the experimental groups and to collect and process data.  
   If no randomization was used, state so.  
   Where does this appear (section, paragraph #)?

   N/A [No treatment assignment]

5. Is a statement of the extent to which investigator knew the group allocation during the experiment and in assessing outcome included?  
   If no blinding was done, state so.  
   Where (section, paragraph #)?

   N/A [No treatment assignment]

6. For experiments in live vertebrates, is a statement of compliance with ethical guidelines/regulations included?  
   Where (section, paragraph #)?

   Yes, page 16 (methods), last paragraph.

7. Is the species of the animals used reported?  
   Where (section, paragraph #)?

   Yes, page 16 (methods), last paragraph.

8. Is the strain of the animals (including background strains of KO/transgenic animals used) reported?  
   Where (section, paragraph #)?

   N/A (no transgenics or special strains were used)

9. Is the sex of the animals/subjects used reported?  
   Where (section, paragraph #)?

   Yes, page 16 (methods), last paragraph.
10. Is the age of the animals/subjects reported?
   Where (section, paragraph #)?
   No.

11. For animals housed in a vivarium, is the light/dark cycle reported?
   Where (section, paragraph #)?
   No.

12. For animals housed in a vivarium, is the housing group (i.e. number of animals per cage) reported?
   Where (section, paragraph #)?
   No, but animals were housed in accordance with National Institute of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

13. For behavioral experiments, is the time of day reported (e.g. light or dark cycle)?
   Where (section, paragraph #)?
   No.

14. Is the previous history of the animals/subjects (e.g. prior drug administration, surgery, behavioral testing) reported?
   Where (section, paragraph #)?
   No.

   a. If multiple behavioral tests were conducted in the same group of animals, is this reported?
      Where (section, paragraph #)?
      N/A

15. If any animals/subjects were excluded from analysis, is this reported?
   Where (section, paragraph #)?
   No animals were excluded.

   a. How were the criteria for exclusion defined?
      Where is this described (section, paragraph #)?
      N/A

   b. Specify reasons for any discrepancy between the number of animals at the beginning and end of the study.
      Where is this described (section, paragraph #)?
      N/A

Reagents

1. Have antibodies been validated for use in the system under study (assay and species)?
   N/A

   a. Is antibody catalog number given?
      Where does this appear (section, paragraph #)?
      N/A
b. Where were the validation data reported (citation, supplementary information, Antibodypedia)?
Where does this appear (section, paragraph #)?

N/A

2. Cell line identity
a. Are any cell lines used in this paper listed in the database of commonly misidentified cell lines maintained by ICLAC and NCBI Biosample?
Where (section, paragraph #)?

N/A

b. If yes, include in the Methods section a scientific justification of their use--indicate here in which section and paragraph the justification can be found.

N/A

c. For each cell line, include in the Methods section a statement that specifies:
- the source of the cell lines
- have the cell lines been authenticated? If so, by which method?
- have the cell lines been tested for mycoplasma contamination?
Where (section, paragraph #)?

N/A

Data deposition

Data deposition in a public repository is mandatory for:
  a. Protein, DNA and RNA sequences
  b. Macromolecular structures
  c. Crystallographic data for small molecules
  d. Microarray data

Deposition is strongly recommended for many other datasets for which structured public repositories exist; more details on our data policy are available here. We encourage the provision of other source data in supplementary information or in unstructured repositories such as Figshare and Dryad.

We encourage publication of Data Descriptors (see Scientific Data) to maximize data reuse.

1. Are accession codes for deposit dates provided?
Where (section, paragraph #)?

N/A

Computer code/software

Any custom algorithm/software that is central to the methods must be supplied by the authors in a usable and readable form for readers at the time of publication. However, referees may ask for this information at any time during the review process.

1. Identify all custom software or scripts that were required to conduct the study and where in the procedures each was used.

N/A
2. If computer code was used to generate results that are central to the paper’s conclusions, include a statement in the Methods section under “Code availability” to indicate whether and how the code can be accessed. Include version information as necessary and any restrictions on availability.

Human subjects

1. Which IRB approved the protocol?
   Where is this stated (section, paragraph #)?
   MIT Committee on the Use of Humans as Experimental Subjects. (Methods, "Human Psychophysical Experiments", p. 19)

2. Is demographic information on all subjects provided?
   Where (section, paragraph #)?
   N/A. The main human data were collected from web-based Amazon Mechanical Turk Platform; subjects’ demographic information was protected by the platform.

3. Is the number of human subjects, their age and sex clearly defined?
   Where (section, paragraph #)?
   The number of subjects is clearly defined (Methods, "Human Psychophysical Experiments", p. 19). Age and sex information was protected by the Amazon MTurk platform and was thus not accessible.

4. Are the inclusion and exclusion criteria (if any) clearly specified?
   Where (section, paragraph #)?
   Yes, as described in answer to Question 3 above (see answer Part b.)

5. How well were the groups matched?
   Where is this information described (section, paragraph #)?
   N/A (no group comparisons)

6. Is a statement included confirming that informed consent was obtained from all subjects?
   Where (section, paragraph #)?
   Confirmation is implied by the fact that our human testing was done in accordance with MIT Committee on the Use of Humans as Experimental Subjects which require consent for all subjects. Methods: Human psychophysics, pg 19.

7. For publication of patient photos, is a statement included confirming that consent to publish was obtained?
   Where (section, paragraph #)?
   N/A

fMRI studies

For papers reporting functional imaging (fMRI) results please ensure that these minimal reporting guidelines are met and that all this information is clearly provided in the methods:

1. Were any subjects scanned but then rejected for the analysis after the data was collected?
   N/A

   a. If yes, is the number rejected and reasons for rejection described?
   N/A

   Where (section, paragraph #)?
2. Is the number of blocks, trials or experimental units per session and/or subjects specified?  
   Where (section, paragraph #)?  
   N/A

3. Is the length of each trial and interval between trials specified?  
   N/A

4. Is a blocked, event-related, or mixed design being used? If applicable, please specify the block length or how the event-related or mixed design was optimized.  
   N/A

5. Is the task design clearly described?  
   Where (section, paragraph #)?  
   N/A

6. How was behavioral performance measured?  
   N/A

7. Is an ANOVA or factorial design being used?  
   N/A

8. For data acquisition, is a whole brain scan used?  
   If not, state area of acquisition.  
   a. How was this region determined?  
      N/A

9. Is the field strength (in Tesla) of the MRI system stated?  
   a. Is the pulse sequence type (gradient/spin echo, EPI/spiral) stated?  
      N/A
   b. Are the field-of-view, matrix size, slice thickness, and TE/TR/flip angle clearly stated?  
      N/A

10. Are the software and specific parameters (model/functions, smoothing kernel size if applicable, etc.) used for data processing and pre-processing clearly stated?  
    N/A

11. Is the coordinate space for the anatomical/functional imaging data clearly defined as subject/native space or standardized stereotaxic space, e.g., original Talairach, MNI305, ICBM152, etc? Where (section, paragraph #)?  
    N/A

12. If there was data normalization/standardization to a specific space template, are the type of transformation (linear vs. nonlinear) used and image types being transformed clearly described? Where (section, paragraph #)?  
    N/A

13. How were anatomical locations determined, e.g., via an automated labeling algorithm (AAL), standardized coordinate database (Talairach daemon), probabilistic atlases, etc.?  
    N/A
14. Were any additional regressors (behavioral covariates, motion etc) used?  | N/A
15. Is the contrast construction clearly defined?  | N/A
16. Is a mixed/random effects or fixed inference used?  | N/A
  a. If fixed effects inference used, is this justified?  | N/A
17. Were repeated measures used (multiple measurements per subject)?  | N/A
  a. If so, are the method to account for within subject correlation and the assumptions made about variance clearly stated?  | N/A
18. If the threshold used for inference and visualization in figures varies, is this clearly stated?  | N/A
19. Are statistical inferences corrected for multiple comparisons?  | N/A
  a. If not, is this labeled as uncorrected?  | N/A
20. Are the results based on an ROI (region of interest) analysis?  | N/A
  a. If so, is the rationale clearly described?  | N/A
  b. How were the ROI’s defined (functional vs anatomical localization)?  | N/A
21. Is there correction for multiple comparisons within each voxel?  | N/A
22. For cluster-wise significance, is the cluster-defining threshold and the corrected significance level defined?  | N/A

Additional comments

Additional Comments

Relatively few statistical tests were done in our study, mainly because our result clearly showed the gap between IT information decoding performance and human consistency confidence values, compared to those from other areas, based on their confidence intervals.