Systems/Circuits

Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance

Najib J. Majaj,[1,2]* Ha Hong,[1,2,3]* Ethan A. Solomon,[1,2] and James J. DiCarlo[1,2]

[1]Department of Brain and Cognitive Sciences, [2]McGovern Institute for Brain Research, and [3]Harvard–Massachusetts Institute of Technology Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

To go beyond qualitative models of the biological substrate of object recognition, we ask: can a single ventral stream neuronal linking hypothesis quantitatively account for core object recognition performance over a broad range of tasks? We measured human performance in 64 object recognition tests using thousands of challenging images that explore shape similarity and identity preserving object variation. We then used multielectrode arrays to measure neuronal population responses to those same images in visual areas V4 and inferior temporal (IT) cortex of monkeys and simulated V1 population responses. We tested leading candidate linking hypotheses and control hypotheses, each postulating how ventral stream neuronal responses underlie object recognition behavior. Specifically, for each hypothesis, we computed the predicted performance on the 64 tests and compared it with the measured pattern of human performance. All tested hypotheses based on low- and mid-level visually evoked activity (pixels, V1, and V4) were very poor predictors of the human behavioral pattern. However, simple learned weighted sums of distributed average IT firing rates exactly predicted the behavioral pattern. More elaborate linking hypotheses relying on IT trial-by-trial correlational structure, finer IT temporal codes, or ones that strictly respect the known spatial substructures of IT ("face patches") did not improve predictive power. Although these results do not reject those more elaborate hypotheses, they suggest a simple, sufficient quantitative model: each object recognition task is learned from the spatially distributed mean firing rates (100 ms) of ~60,000 IT neurons and is executed as a simple weighted sum of those firing rates.

Key words: categorization; identification; invariance; IT cortex; object recognition; V4

Significance Statement

We sought to go beyond qualitative models of visual object recognition and determine whether a single neuronal linking hypothesis can quantitatively account for core object recognition behavior. To achieve this, we designed a database of images for evaluating object recognition performance. We used multielectrode arrays to characterize hundreds of neurons in the visual ventral stream of nonhuman primates and measured the object recognition performance of >100 human observers. Remarkably, we found that simple learned weighted sums of firing rates of neurons in monkey inferior temporal (IT) cortex accurately predicted human performance. Although previous work led us to expect that IT would outperform V4, we were surprised by the quantitative precision with which simple IT-based linking hypotheses accounted for human behavior.

Introduction

The detailed mechanisms of how the brain accomplishes viewpoint invariant object recognition remain largely unknown, but lesion studies point to the ventral stream [V1–V2–V4–inferior temporal (IT) cortex] as being critical to this behavior (Holmes and Gross, 1984; Biederman et al., 1997). Previous ventral stream studies have focused on understanding the nonlinear transformations between the retina and neural responses (Gallant et al., 1996; Hegdé and Van Essen, 2000; Pasupathy and Connor, 2002; Connor et al., 2007; Freeman et al., 2013), including evidence that IT is better at recognition than early representations (Hung et al.,

Received Dec. 20, 2014; revised July 10, 2015; accepted Aug. 24, 2015.

Author contributions: N.J.M., H.H., and J.J.D. designed research; N.J.M., H.H., E.A.S., and J.J.D. performed research; N.J.M., H.H., E.A.S., and J.J.D. analyzed data; N.J.M., H.H., and J.J.D. wrote the paper.

This work was supported by Defense Advanced Research Projects Agency (DARPA Neovision2), the National Institutes of Health (Grant NEI-R01 EY014970 to J.J.D.), and the National Science Foundation (Grant IIS-0964269 to J.J.D.). H.H. was supported by a fellowship from the Samsung Scholarship. We thank Nicolas Pinto for help in creating the framework to generate the images and help with the design and testing of the computer vision models.

The authors declare no competing financial interests.

*N.J.M. and H.H. contributed equally to this work.

Correspondence should be addressed to Najib J. Majaj, Center for Neural Science, New York University, 4 Washington Place, Room 809, New York, NY 10003. E-mail: najib.majaj@nyu.edu.

DOI:10.1523/JNEUROSCI.5181-14.2015

Copyright © 2015 the authors 0270-6474/15/3513402-17$15.00/0

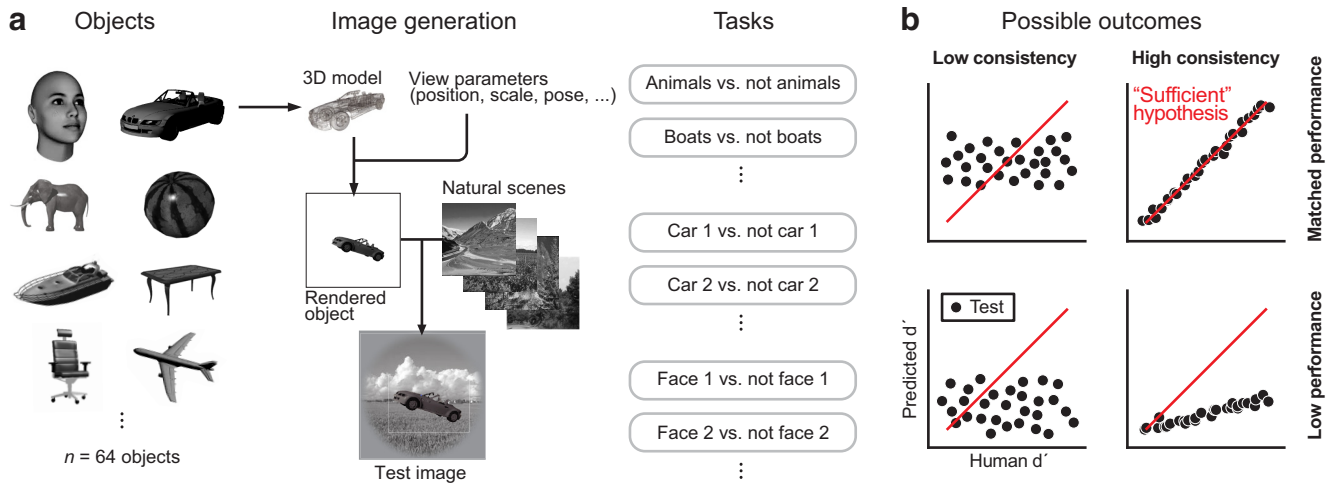

Figure 1. ***a***, Object recognition tasks. To explore a natural distribution of shape similarity, we started with eight basic-level object categories and picked eight exemplars per category resulting in a database of 64 3D object models. To explore identity preserving image variation, we used ray-tracing algorithms to render 2D images of the 3D models while varying position, size, and pose concomitantly. In each image, six parameters (horizontal and vertical position, size, rotation around the three cardinal axes) were randomly picked from predetermined ranges (see Materials and Methods). The object was then added to a randomly chosen background. All test images were achromatic. Human observers performed all tasks using an 8-way approach (i.e., see one image, choose among eight; see Materials and Methods). Two kinds of object recognition tasks were tested: basic-level categorization (e.g., "car" vs "not car") and subordinate identification (e.g., "car 1" vs "not car 1"). We characterized performance for each of eight binary tasks (e.g., "animals" vs "not animals," "boats" vs "not boats," etc.) in each 8-way recognition block at two to three levels of variation, resulting in 64 behavioral tests (64 d' values). ***b***, Possible outcomes for each tested linking hypothesis. We defined multiple candidate neuronal and computational linking hypotheses (Fig. 5), determined the predicted (i.e., cross-validated) object recognition accuracy (d') of each linking hypothesis on the same 64 tests (y-axis in each scatter plot), and compared those results with the measured human d' (x-axis in each scatter plot). A priori, each tested linking hypothesis could produce at least four possible types of outcomes. The pattern of predicted d' might be unrelated to or strongly related to human d' (left vs right scatter plots). We quantified that by computing consistency, the correlation between predicted d' and actual human d' across all 64 object recognition tests. Average predicted d' might be low or matched to human d' (bottom vs top scatter plots). We quantified performance by computing the median ratio of predicted d' and actual human d', across all 64 object recognition tests. For brevity, we will refer to these two metrics as consistency and performance from here on. Our goal was to find at least one "sufficient" code: a linking hypothesis that perfectly predicted the human d' results on all object recognition tests (top right scatter plot).

2005; Rust and DiCarlo, 2010) and that IT responses are partially correlated with perceptual report (Sheinberg and Logothetis, 1997; Op de Beeck et al., 2001; Kriegeskorte et al., 2008). Although such studies tell us much about visual processing and support the belief that the ventral stream is critical to object representation, they do not present a single linking hypothesis that is *quantitatively* sufficient to explain how ventral stream neural activity accounts for object recognition performance over all recognition tests. This study aimed to provide that link for a subdomain of object recognition, core object recognition (DiCarlo et al., 2012), in which images are presented for 100 ms in the central 10° of the visual field.

Our strategy was as follows: (1) develop a stringent behavioral assay, (2) obtain sufficient neuronal sampling, (3) implement specific hypotheses that each predict perceptual report from neural activity, and (4) compare those predictions with actual perceptual reports. We addressed each challenge as follows.

We characterized human core object recognition performance using large image sets that explore shape similarity and identity preserving image variation and assumed that monkey and human patterns of performance are equivalent (see Discussion). The 64 recognition tests that we used set a high bar because performance on them varies widely and is not explained by low-level visual representations.

We measured neuronal responses in visual area V4 and along IT cortex (Felleman and Van Essen, 1991) using the same pool of images used in the behavioral testing. We relied on multielectrode arrays to monitor hundreds of sites, each tested with multiple repeats of 5760 images. Our measured neuronal population was adequate for quantifying uncertainty with respect to neuronal sampling and allowed us to extrapolate to larger numbers of neurons.

We tested specific quantitative versions of previously proposed hypothetical links between neuronal activity and recognition behavior, as well as control hypotheses. Each neuronal linking hypothesis is a postulated mechanism of how downstream neurons integrate ventral stream activity to make a decision about which object label the observer will report in each image (Connor et al., 1990; Parker and Newsome, 1998; Johnson et al., 2002).

Ideally, a sufficient linking hypothesis should predict the perceptual report for each and every image. Here, we focused on predicting the mean human recognition accuracy (d') for all object recognition tests, with each test containing many images. Specifically, we compared the predicted pattern of d's with that measured in humans on the same 64 tests (which could lead to a range of outcomes; Fig. 1*b*).

We report here that simple, learned weighted sums of randomly selected average neuronal responses spatially distributed over monkey IT (referred to here as "LaWS of RAD IT") are able to meet that high bar (Fig. 1*b*, top right). In contrast, other linking hypotheses based on neuronal responses from IT or other visual areas fall short. Although this is compatible with previous ideas about IT's role in object recognition (Tanaka, 1993; Kobatake and Tanaka, 1994; Tanaka, 1997), it is, to our knowledge, the first demonstration that a single, specific neural linking hypothesis is quantitatively sufficient to explain behavior over a wide range of core object recognition tasks.

Materials and Methods

Sixty-four-object recognition tests

To characterize human object recognition abilities (which we assume are similar to those of monkeys; see Discussion), we designed a behavioral assessment tool with images and tasks that span the range of human performance in core object recognition. To explore shape similarity, we

used objects that can be parsed into basic-level categories with multiple exemplars per category, allowing us to test human performance on coarse and fine discriminations. To explore identity preserving object transformations, the "invariance problem," a hallmark of object recognition (DiCarlo and Cox, 2007; DiCarlo et al., 2012), we used ray-tracing software to photorealistically render each object while parametrically varying its position, size, and pose. Finally, to insure that the tasks were challenging for current computer vision algorithms, we placed each object on a randomly chosen natural background that was uncorrelated with its identity (Pinto et al., 2008). To focus on the so-called "core object recognition," that is, recognition during a single, natural viewing fixation (DiCarlo et al., 2012), each test image was presented as an 8° patch directly at the center of gaze for 100 ms. The culmination of our effort was a set of 64 core object recognition tests (24 noun labels, each at 2 or 3 levels of variation; see Fig. 3) that constitutes a reasoned attempt at exploring the power of human object recognition. We do not claim this to be an exhaustive characterization of human object recognition, but rather an initial operational definition that can be sharpened and extended to explore other aspects of object recognition and shape discrimination (see Discussion).

Image generation

High-quality images of single objects were generated using free ray-tracing software (http://www.povray.org). Each image consisted of a 2D projection of a 3D model (purchased from Dosch Design and TurboSquid) added to a random background. No two images had the same background. In a few cases, the background was by chance correlated with the object (plane on a sky background), but mostly they were uncorrelated, with the background on average giving no information about the identity of the object.

This general approach allowed us to create a database of 5760 images based on 64 objects. The objects were chosen based on eight "basic-level" categories (animals, boats, cars, chairs, faces, fruits, planes, tables), with eight exemplars per category (BMW, Z3, Ford, etc.). By varying six viewing parameters, we explored three types of identity while preserving object variation, position (x and y), rotation (x, y, and z), and size. The parameters were varied concomitantly and each was picked randomly from a uniform range that corresponded to one of three levels of variation (low, medium, and high). For the low variation image set, the parameters were fixed and picked to correspond to a fixed view and size of each object centered on the background. For example, cars were presented in their side view, whereas faces were presented with a frontal view. However, we did vary the backgrounds so that each object was presented on 10 randomly picked backgrounds, resulting in a total of 640 images. For medium and high variation, we generated 40 images per object, resulting in 2560 images per variation (total $5760 = 640 + 2 \times 2560$). Each image was rendered with a pooled random sample of the six parameters and presented on a randomly picked background. The parameters of medium variation had the following ranges for x- and y-position, pose (all 3 axes treated independently), and size: [−1.2°, 1.2°], [−2.4°, 2.4°], [−45°, 45°], and [×1/1.3, ×1.3]. Those for high variation were [−2.4°, 2.4°], [−4.8°, 4.8°], [−90°, 90°], and [×1/1.6, ×1.6]. All images were achromatic with a native resolution of 256 × 256 pixels (see Fig. 3 for example images).

Human psychophysics and analysis

All human studies were done in accordance with the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects. A total of 104 observers participated in 1 of 3 visual task sets: an 8-way classification of images of 8 different cars, an 8-way classification of images of 8 different faces, or an 8-way categorization of images of objects from 8 different basic-level categories. Observers completed these 30–45 min tasks through Amazon's Mechanical Turk, an online platform in which subjects can complete experiments for a small payment. All of the results were confirmed in the laboratory setting with controlled viewing conditions and virtually identical results were obtained in the laboratory and online populations (Pearson correlation = 0.94 ± 0.01).

Each trial started with a central fixation point that lasted for 500 ms, after which an image appeared at the center of the screen for 100 ms. After a 300 ms delay, the observer was prompted to click one of eight "response" images that matched the identity or category of the stimulus image. The image presentation time was chosen based on results showing that core object recognition performance improves quickly over time such that accuracy for a 100 ms presentation time is within 92% of performance at 2 s (see Fig. S2 in Cadieu et al., 2014). Results were very similar, with slightly shorter (50 ms) or longer (200 ms) viewing duration. To enforce the need for view tolerant "object" recognition (rather than image matching), each response image displayed a single object from a canonical view without background. After clicking a response image, the subject was given another fixation point before the next stimulus appeared. No feedback was given. The "response" images remained constant throughout a block of trials that corresponded to one set of tasks (i.e., an 8-way categorization block contained eight embedded binary tasks).

Human object recognition performance was determined by computing a d' for each binary task. Specifically, for a given 8-way task set and variation level (e.g., basic-level categorization at hard variation or car subordinate identification at medium variation), we constructed the raw 8×8 confusion matrix for each individual observer. Then, we computed the population confusion matrix by taking the sum of these raw confusion matrices across individuals. From the population confusion matrix, we computed the d' for each task of recognizing one target class against seven distractor classes (i.e., the "binary" task). We obtained 72 d' measurements by performing this procedure over all combinations of 3 task sets and 3 variation levels (3 task sets × 8 targets per task set × 3 levels of variation). We excluded face identification at high variation because none of the eight d's were statistically distinguishable from random guessing, leaving a total of 64 behavioral tests for the main results presented. Inclusion of these eight d's had no significant effect on the results.

Typically, each human observer only participated in one of the three test sets (basic-level categorization and car and face subordinate-level identification test set; four of 104 subjects participated in both the car and face test sets). For the 8-way basic-level categorization test set, each observer ($n = 29$) judged a subset of 400 randomly sampled images at each variation level (400 of 640 for low variation and 400 of 2560 for medium and high variation levels). For the 8-way car ($n = 39$) and 8-way face ($n = 40$) identification test sets, each observer saw all 80 images at the low variation level and all 320 images at both medium and high variation levels. The presentation of images was randomized and counterbalanced so that the number of presentations of each class was approximately the same in a given variation level. Variation levels were presented in successively harder blocks. Observers would see a full set of low variation ("easy") images before moving to medium and then high variation ("difficult") images. On a few additional observers ($n = 10$), we interleaved the different test sets (basic categorization, car and face identification at low variation) and saw no significant effect of interleaving on the pooled population d's (the Pearson correlation coefficient between blocked and interleaved was 0.903 ± 0.057; see the next paragraph for the procedures to compute the pooled population d's).

Although no single observer judged all the images in our image database, our pool of human observers did. To compute the pooled population human d's, we started with each observer's data and computed a 8×8 confusion matrix for each variation level. We then constructed the population-confusion matrix for each test set and variation level (e.g., 8-way low variation car identification) by summing across individual subject's confusion matrices. We used standard signal detection theory to compute population d's from the pooled population confusion matrix ($d' = Z(\text{TPR}) - Z(\text{FPR})$, where Z is the inverse of the cumulative Gaussian distribution function and TPR and FPR are true-positive and false-positive rates, respectively).

The 64 human population d's were the benchmark to which we compared our candidate linking hypotheses. To capture the four possibilities for such a comparison (Fig. 1*b*), we defined two metrics, *consistency* and relative *performance*. To quantify the match between the pattern of predicted and human d's, we computed consistency, the rank correlation between predicted d' and actual human d' across all 64 object recognition tests. To quantify the match on average between predicted and actual human d', we computed relative performance, the median ratio of pre-

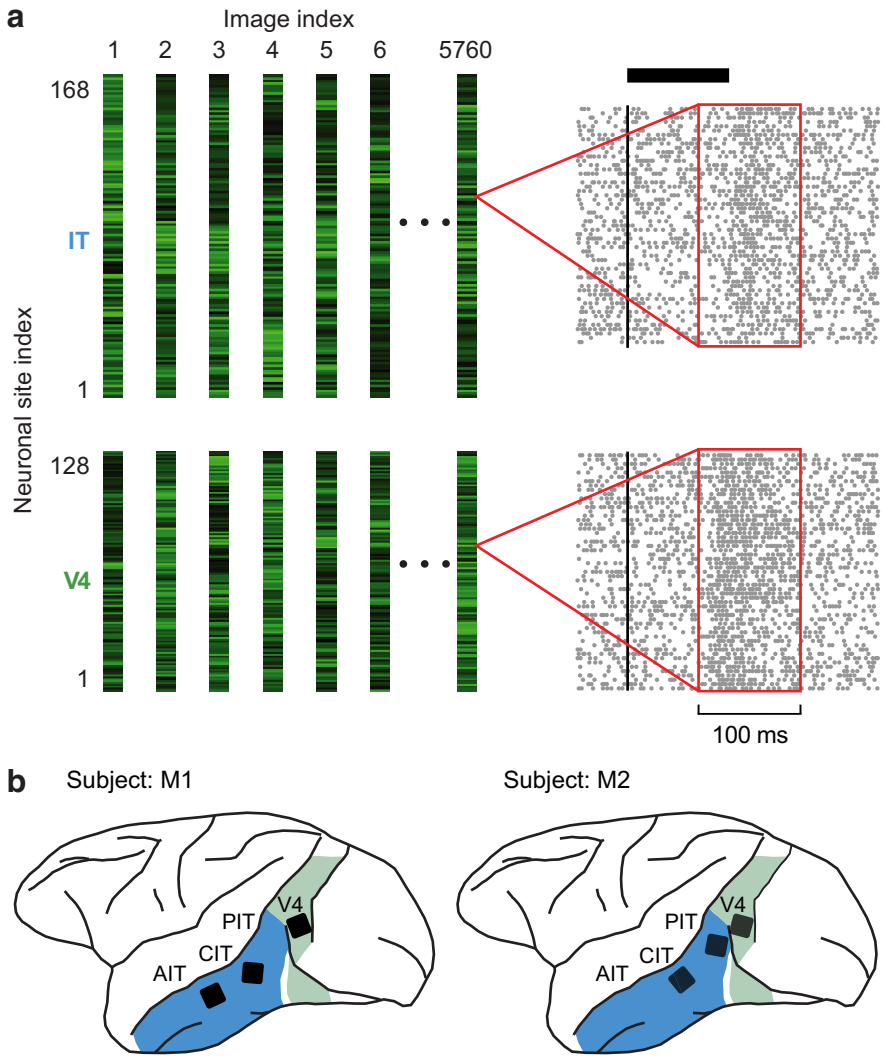

Figure 2. Neural responses. ***a***, We used multielectrode arrays to record neural activity from two stages of the ventral visual stream [V4 and IT (PIT + CIT + AIT)] of alert rhesus macaque monkeys. We recorded neural responses to the same images used in our human psychophysical testing. Each image was presented multiple times (typically ~50 repeats, minimum 28) using standard rapid visual stimulus presentation (RSVP). Each stimulus was presented for 100 ms (black horizontal bar) with 100 ms of neutral gray background interleaved between images. Although some of our neural sites represented single neurons, the majority of our responses were multiunit (see Fig. 8*a*). The rasters for repeated image presentations were then tallied within a defined time window (e.g., 70–170 ms after image onset, red rectangle, black vertical line indicated stimulus onset) to compute an average firing rate in impulses per second. The mean evoked firing rate is an entry in a response vector (green vertical vector, green saturation is proportional to response magnitude) that summarizes the population response to a single image. The concatenation of the response vectors produces a response matrix representing the population neural response of a particular visual area to our database of 5760 images. We parsed our neural population into V4 and IT, treating the various parts of IT as one population. We recorded from 168 neural sites in IT and 128 neural sites in V4. ***b***, Approximate placement of the arrays in V4 (green shaded areas) and IT (blue shaded area) is illustrated by the black squares on two line drawings representing the brains of our two subjects.

dicted d', and actual human d' across all 64 object recognition tests. To estimate the human subject-to-subject variability for consistency and performance, we selected one subject from each test set and combined the test performance of the three test sets to produce 64 "individual" human d's (Fig. 3*b*). We repeated this procedure multiple times to construct an ensemble of individual human subjects. We used this ensemble to compute the median consistency (Spearman rank correlation coefficient) and performance (1 by definition) between individual human d's and pooled population human d's and the 68% confidence intervals around that median.

To investigate the effect of image subsampling on our results, we computed the sampling induced SE of the pooled population d's on the basic-level categorization test set. The SE was minimal (median = 2.1% of corresponding d') because the entire image set was presented multiple times to our large pool of observers ($n = 29$). Assuming the effect of this error to be additive and independent, the predicted consistencies of a linking hypothesis would be increased by ~0.15% if each of our observers judged the entire 5760 image in the basic-level categorization test set.

We also generated two predictions on how consistency might improve if we had collected human data on all images. If we assume that the human-to-human consistency will eventually be 1 as the number of presented images increases to infinity, the Spearman–Brown prediction formula allows us to estimate the human-to-human consistency and its confidence interval (CI) as if we had collected human data on all images in our image set. This assumption resulted in an increase of only ~1.9% to the human-to-human consistency and the CI results presented in the main text. If we assumed a more reasonable asymptote of 0.95, the increase was only ~0.59%. In combination, the above two analyses suggest that image subsampling in the human basic-level categorization test set had no significant effect on our main results.

Animals, surgeries, and training

The nonhuman subjects in this experiment were two adult male rhesus monkeys (*Macaca mulatta*, 7 and 9 kg). Before training, we surgically implanted each monkey with a head post under aseptic conditions. We monitored eye movement using video eye tracking (SR Research EyeLink II). Using operant conditioning and juice reward, our 2 subjects were trained to fixate a central red square (0.25°) within a square fixation window that ranged from ±1° to ±2.5° for up to 4 s. Outside of maintaining fixation, no additional attempt was made at controlling spatial or feature attention.

We recorded neural activity using 10 × 10 microelectrode arrays (Blackrock Microsystems). A total of 96 electrodes were connected; the corners were not connected. Each electrode was 1.5 mm long and the distance between adjacent electrodes was 400 μm. Before recording, we implanted each monkey with three arrays in the left cerebral hemisphere, one array in V4, and two arrays in IT, as shown in Figure 2*b*. Array placement was guided by the sulcus pattern, which was visible during surgery. The electrodes were accessed through a percutaneous connector that allowed simultaneous recording from all 96 electrodes from each array (three connectors on each animal). All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and real-time video eye tracking. All surgical and animal procedures were performed in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

Monkey behavior, image presentation, and recording procedures

Our goal was to assess neuronal activity patterns that are automatically evoked by presenting a visual image to an awake, alert visual system. Therefore, the monkey's only task was to maintain gaze on a fixation dot in the middle of a screen for 2–4 s as images were serially presented at the center of gaze. The monkeys initiated each trial by fixating a central red square. After initiating fixation (160 ms), a sequence of 5–10 images was

presented for 100 ms each with 100 ms of blank screen in between. Each image was presented at the center of gaze and subtended 8° of the visual field with a resolution of 32 pixels/° and a pixel luminance range of 0.3–300 cd/m^2. The images were presented in a randomized order and each image was presented for at least 28 repetitions (typically ~50). We recorded neural responses for 5760 images drawn from the same pool that we used in our human psychophysical testing, with nearly identical visual presentation parameters.

During each recording session, band-pass filtered (250 Hz to 7.5 kHz) neural activity was monitored continuously and sampled at 30 kHz using commercially available hardware (Blackrock Microsystems). The majority of the data presented here were based on multiunit activity [MUA; see Fig. 8*a* for single unit activity (SUA) analysis]. A multiunit spike event was defined as the threshold crossing when voltage (falling edge) deviated by less than three times the root mean squared error of baseline voltage. Threshold was typically set once during the beginning of a recording session while the animal was viewing a blank gray screen. Of 576 implanted electrodes (three arrays × 96 electrodes × two monkeys), we focused on the 296 [128 V4 and 168 across posterior (PIT), central (CIT), and anterior (AIT) inferior temporal cortex)] most visually driven neural sites. To pick these sites, we estimated evoked visual response using an independent set of images (typically 795 images with a minimum of 350 images). Visual drive was then defined as the cross-validated average of the top 10% evoked image responses (d' between neural response to image versus blank). Receptive fields were mapped with briefly flashed bars and the expected contralateral receptive field biases were observed in V4.

We recorded all spike time events at all neural sites. As described in the text, we defined different neuronal codes by considering spike counts in different time windows relative to image presentation. Our array placements allowed us to sample neural sites from V4 and different parts of IT. For most analyses, we grouped all sites into either a V4 or an IT population. The response of all neural sites in a population (V4 or IT) to an image formed a vector; the image vectors in turn formed a matrix (described further below) summarizing the population response to all 5760 tested images (Fig. 2*a*). To fill a response matrix and its multiple repetitions, neural responses were collected over multiple days [68 d for Monkey 1 (M1) and 65 d for Monkey 2 (M2); stability and its impact on the results is discussed below].

Construction of specific, candidate-linking hypotheses and their predicted behavioral performance

A neuronal linking hypothesis is a formal rule for converting neural activity to overt behavior (e.g., a choice of object class/label). Here, each candidate-linking hypothesis learns a neural code that converts neural responses into a prediction of the type of object that is present in the world (as conveyed by the visual image). Defining each linking hypothesis requires the specification of two components. The first is a "response matrix" of neural (or, in some cases, computer-generated) responses to each image. This specification includes which neurons are included (e.g., responses of 100 spatially distributed IT neurons), as well as a specification of the relevant aspects of that neural activity (e.g., time window, mean rate). The second component is a specific type of presumed downstream neural decoder, along with a training procedure for the decoder that specifies how to estimate its final learned state. After specifying these two components for each linking hypothesis, we computed its predicted behavioral performance for each of the 64 object recognition tests using independent test images.

Response matrix

Neural response matrix. For neuronal linking hypotheses, the response matrix is a $N \times I$ matrix in which N is the number of neuronal sites considered to be part of the linking hypothesis and I is the total number of images tested. Because our image set was very large ($N = 5760$), we collected neural responses piecemeal over multiple days. Each entry of the matrix is the "response" of a particular neural site to a particular image. We considered V4 and IT separately. For each visual area, the "response" was computed as follows. First, we counted the number of spike events elicited by each image in each neural site over a given time window. For example, one possibility is the time window 70–170 ms after image onset, but many other possibilities exist and we explored some of those. From this response, we subtracted the neural site's background response for that day (mean response to "blank" images). Finally, the evoked response of each neural site was normalized by the site's sample SD (over all tested images that day). This normalization was done to compensate for day-to-day variation and had no effect on pattern of performance and a small effect on absolute performance (~5% increment). The full matrix was collected multiple times (typically ~50 repetitions, minimum 28) and averaged across all repetitions.

Feature response matrix. We also constructed linking hypotheses in which the "responses" were simulated rather than being directly measured from neural activity. These included pixels, V1-like model neurons, and several popular algorithms in the computer vision community. These algorithms each take an image and produce the values of a fixed number of "features" (operators on the image). For each algorithm, we computed the response of all of its feature outputs for each of our images. We treated these feature outputs as being analogous to neuronal populations and thus constructed a response matrix for each algorithm. We explored pixel (n = ~16 k features that have comparable visual drivenness as neuronal features), V1-like (n = ~76 k features, again, visually driven), PHOG (n = ~3 k visually driven features), SIFT (n = ~59 k visually driven features), an HMAX variant called sparse localized features (SLFs; n = ~4 k visually driven features), and an L3 algorithm (n = ~4 k visually driven features) (Pinto et al., 2011) (details discussed below).

Downstream neuronal decoders and training procedures

To estimate what information downstream neurons could easily "read" from a given neural population, we used simple, biologically plausible linear decoders (i.e., linear classifiers, linear discriminants). Such decoders are simple in that they can perform binary classifications by computing weighted sums (each weight is analogous to the strength of synapse) of input features and separate the outputs based on a decision boundary (analogous to a neuron's spiking threshold). The decoders differ in how the optimal weights and decision boundary are learned. We mainly explored two types of linear decoders, support vector machines (SVMs) and correlation-based decoders (CCs). The SVM learning model generates a decoder with a decision boundary that is optimized to best separate images of the target object from images of the distractor objects. The optimization is done under a regularization constraint that limits the complexity of the boundary. We used LibSVM software package (Chang and Lin, 2011) with the linear C-SVC algorithm and L_2 regularization (the regularization constant C was set to 5×10^4 except for the linking hypotheses in Fig. 5, *c–e*, where the C was optimized by a 3-fold cross-validation on training data). The CC learning model (Meyers et al., 2008) produces a decoder using the target class center estimated by computing the mean across the target images in the training data. The resulting decoder determines the test image's membership by computing the Pearson correlation coefficient between the target class center and the image. Correlation-based decoders are simpler than SVMs in two regards: (1) they are determined by class centers in the training data without mathematical optimization and (2) they do not have free parameters that are unrelated to the data that affect the optimization procedure (Meyers et al., 2008). For completeness, we also explored simpler single feature decoders (max, 95th quantile, 90th quantile, median). These decoders were built by searching for features based on certain criteria. For example, a "max" decoder is built by finding the feature or neural site that has the best d' for each behavioral test. All of these decoders could potentially be implemented by downstream neurons because they involve two basic operations: weighted sums of inputs followed by a threshold.

For a given task set (e.g., 8-way basic-level classification) and variation level (low, medium, or high) (see "Human psychophysics" section above for details), the corresponding portion of the response matrix was split into "training" and "testing" sets. The mean and variance of each unit or feature was normalized so that its responses to the training set have zero mean and unit variance. The training set was then used to optimize eight "one-vs-rest" linear decoders by finding weights that would maximize classification performance of each. To construct an 8-way decoder, analogous to what the human observers were asked to do, we applied all eight

decoders and scored the decoder with the largest output margin as the predicted behavioral "choice" of the linking hypothesis.

Generating the predicted behavioral performance of each candidate linking hypothesis

After constructing each candidate linking hypothesis (i.e., after learning how to read the "code" for each task), we used the "testing" image set (never seen by the decoder) to generate the linking hypothesis's predicted behavioral performance in each of the three task sets. Each such 8-way classification scheme resulted in an 8×8 confusion matrix summarizing the predicted performance (hits and false arms) of a particular linking hypothesis on a particular task set and variation level. This was done multiple times with at least 50 training/testing splits. The average confusion matrix across all splits was then used to compute predicted d's that are exactly analogous to the measured human d's. We also tested a binary two-way classification scheme more common in the computer vision community. The two alternative schemas resulted in similar absolute performance (~5% difference in average performance level) and practically identical pattern of performance (~2% difference in consistency with humans).

Face selectivity index

We defined face-selective IT sites as the ones that have an absolute face selectivity index (FSI) larger than 1/3. The FSI of a site was computed as follows (Tsao et al., 2006; Issa and DiCarlo, 2012):

$FSI = (F - NF)/(|F| + |NF|)$ where F and NF denote the site's mean response to face and nonface stimuli, respectively.

Stability and assumption of combining neural activity across recording days

To collect a large number of repetitions from the thousands of tested images, we had to collect data from the recording arrays over ~45 d (M1, 43 d; M2, 47 d). Although the recording arrays are fixed in tissue and are thus sampling the same cortical location across days, these methods cannot guarantee that the exact same neurons are recorded over all days. Such absolute stability, although desirable, is not strictly required to test the neuronal linking hypotheses that we consider here (which assume randomly selected samples of IT neurons). Nevertheless, we sought to understand whether our presented results might be different if the exact same neurons had been recorded over all days. To do this, we compared performance obtained by averaging the neural responses to six presentations of all images collected on the same day (assuming stable set of neurons during the day) to performance obtained by averaging the responses to the same number of image presentations ($n = 6$), but sampled randomly from multiple days without replacement (always sampled from the same electrode). Each of the two methods produced a pattern of 64 predicted d' values (as in the main text) and we found that those patterns were very similar—the mean Pearson correlation coefficients between the two sets of performances was 0.908 ($\pm$ SD of 0.016 across different samples of trials; $n = 64$ d's) for IT and 0.923 ($\pm$ SD of 0.016) for V4. Therefore, although it is possible that there is some day-to-day variation of recorded activity on each electrode, that variation is small in that it does not substantially change the pattern of performances (e.g., some IT linking hypotheses predict human performance and V4 linking hypotheses do not) and thus is unlikely to change our main result.

Consistency and performance of neuronal linking hypotheses when objects were presented in the ipsilateral versus the contralateral visual field

Because all arrays were placed in the left hemisphere, we wondered whether performance of our neuronal linking hypotheses was affected by object position in the visual field. To address this, we divided the response matrix of each visual area (V4, IT) into two groups based on whether the object centers were in the ipsilateral or contralateral visual field. We then compared performance on the two groups of images using analogous training and testing procedures to what we used for our main results. Consistent with the known contralateral visual field bias in V4, our results showed an ~20% reduction in performance of V4 for ipsilaterally presented objects, whereas IT showed only an ~3% reduction. However, even when only considering objects in the contralateral visual field, the pattern of behavioral performance predicted by V4 was still very different from the actual human performance (consistency = 0.470 $\pm$ 0.111, error is computed by sampling over of behavioral tests and assuming that human pattern of performance does not depend on visual hemifield).

Consistency and performance of neuronal linking hypotheses when objects were only presented foveally

Because V4 units typically have smaller receptive fields than eccentricity-matched IT units and because the array placements focused on foveal V4, we also wondered whether V4-based linking hypotheses could be improved by restricting our image set to objects positioned close to the fovea. To test this hypothesis, we remeasured human behavior and neuronal responses (V4 and IT) for a new set of images that did not contain any variation in object position. We used a total of 32 behavioral tests—24 low variation tests (eight basic-level, eight car identification, and eight face identification tests) and eight new basic-level tests based on images rendered specifically for this analysis (i.e., objects were rendered with randomly picked pose (rotation around x, y, and z) and size parameters, but position (x, y) was fixed at the center of the image). Each linking hypothesis consisted of 58 units and we used correlation decoders for this analysis. All other details were optimized to obtain best performance. For this set of 32 tests, the median human-to-human consistency was 0.887 (with the 68% CI = [0.740, 0.947] due to the sampling of individuals and object recognition tests). The consistency between the LaWS of RAD IT linking hypothesis and human performance was 0.868 (with a 68% CI = [0.791, 0.909] due to the sampling of behavioral tests). The consistency between the LaWS of RAD V4 linking hypothesis and human performance was −0.196 (CI = [−0.358, 0.001]). Although the performance of the LaWS of RAD IT linking hypothesis was indistinguishable from human subjects in terms of consistency ($p = 0.411$, bootstrap test), the LaWS of RAD V4 linking hypothesis had significantly lower consistency ($p < 0.001$, bootstrap test). This low consistency was not caused by the low performance of the V4-based linking hypotheses (similar to Fig. 7*a*, open green circles); in 12 behavioral tests (usually low variation identification tests), V4-based linking hypotheses outperformed the pooled human population. This analysis confirms that the performance of these V4 linking hypotheses is not limited by receptive field size and argues instead for an inferior and potentially more tangled V4 representation (DiCarlo and Cox, 2007; DiCarlo et al., 2012).

Computer vision algorithm hypotheses considered

We compared our biological results on consistency and performance to a variety of computational models, including: The trivial pixel control, in which the original 256×256 square images were down-sampled into 150×150 pixels and flattened into a 22500-dimensional "feature" representation. The pixel features provided a control against the most basic types of low-level image confounds. All of the following computer vision features were computed based on this downsized 150×150 pixel feature. We used an optimized V1-like model, built on grid of Gabor edges at a variety of frequencies, phases, and orientations (Pinto et al., 2008), with each image represented by 86400 features. We also used PHOG (Pyramid Histogram Of Gradients), a spatial pyramid representation of shape based on orientations gradients of edges extracted with a Canny detector (Lazebnik et al., 2009). We fixed the angular range to 360° and the number of quantization bins to 40 to produce 3400-dimensional features. The baseline SIFT computer vision model provided another control against low-level image confounds (Lowe, 2004). The SIFT descriptors were computed on a uniform dense grid with a spacing of 10 pixels and a single patch size of 32 by 32 pixels. Each image was represented by 67712 features. The bio-inspired SLFs are extensions of the C2 features from the HMAX model (Riesenhuber and Poggio, 1999; Serre et al., 2007; Mutch and Lowe, 2008). HMAX is a multilayer convolutional neural network model targeted at modeling higher ventral cortex. Because it is a deep network, HMAX has large, IT-like receptive fields. HMAX is one of many existing "first-principles"-based models that attempt to build up invariance through hierarchical alternation of simple and complex cell-like layers. There were 4096 features per image. L3 is a recent three-layer convolutional neural network, which also has large IT-like receptive

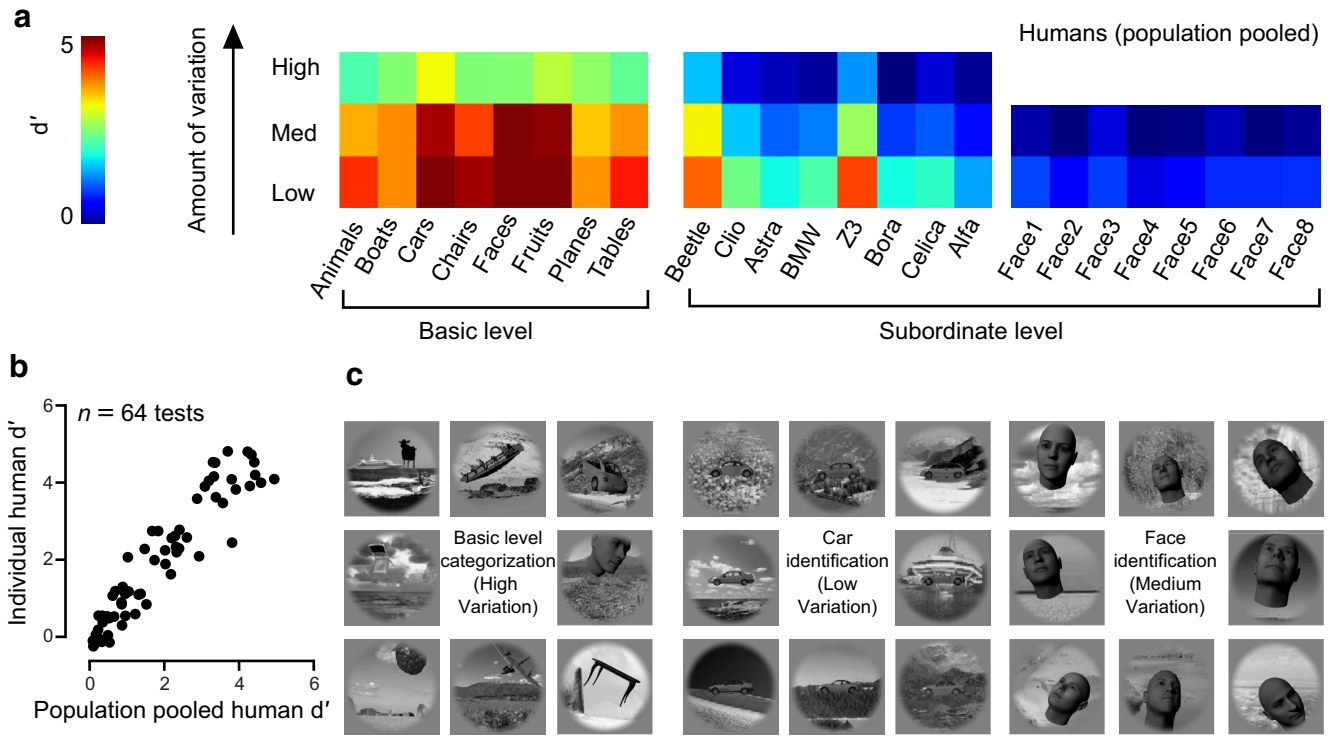

Figure 3. Human core object recognition results. ***a***, Each color matrix from left to right summarizes the pooled human d' for each of the three task sets ranging from basic level categorization to subordinate level face identification. In each matrix, the amount of identity preserving image variation was increased from low (bottom) to high (top), resulting in a total of 64 behavioral tests. Red represents high performance ($d' = 5$) and blue low performance ($d' = 0$). For each 8-way test set and each level of variation, the computed eight d's were based on the average confusion matrix of multiple observers (basic level categorization, $n = 29$, car identification, $n = 39$, face identification, $n = 40$; see Materials and Methods for more information). ***b***, Human to human consistency. The scattergram shows the performance (d') of one human observer plotted against the performance (d') of the pooled population of human observers across all 64 tests. The individual human observer was created by randomly combining the performance of three subjects on the three test sets (basic-level categorization and car and face subordinate-level identification). The population performance was computed based on a confusion matrix that combined the judgment of our entire pool ($n = 104$) of human observers. The Spearman correlation coefficient in this example was 0.941 (with a 68% CI = [0.921, 0.946] over the choice of behavioral tests). Median relative performance was 0.999 (with a test-induced 68% CI = [0.965, 1.073] over the choice of behavioral tests). ***c***, Example images. Each octet of images are image samples representing all eight objects used for each of the three tested task sets at three example variation levels: basic level categorization (high variation), car identification (low variation), face identification (medium variation).

fields and which was discovered via a high-throughput screening procedure (Pinto et al., 2008). We used the top-five models identified in Pinto et al. (2008) and the dimensionality of each was 15488, 6400, 2048, 4608, and 10368, respectively.

Results

Quantitative characterization of human core object recognition

To characterize human core object recognition abilities (DiCarlo et al., 2012), we designed 64 core object recognition tests and obtained an unbiased measure of human ability (d'; see Materials and Methods) on each test. Figure 3*a* uses a color scale to show human d' values for each of the 64 tests. The wide range of values is not due to subject variability because the pattern of values over the 64 tests from any one subject is highly correlated with the pattern of values from the pooled results of all other subjects (median correlation = 0.93; Fig. 3*b*). Instead, it shows that all humans find some object recognition tests to be easy ($d' \sim 5$; corresponds to an unbiased accuracy of 99.4% correct, where 50% is chance), others to be more difficult ($d' \sim 2$; 84.1% correct), and others to be very challenging ($d' \sim 0.5$; 59.9% correct). Two unsurprising qualitative trends are noted. First, human object recognition ability is dependent on shape similarity: we found high d's for basic-level categorization tests ("cars" vs "noncars," "faces" vs "nonfaces," "animals" vs "nonanimals," etc.; mean d' across all levels of object view variation = 3.46), lower d's for car identification tests (easy subordinate, e.g., "car1" vs "not car1"; mean $d' = 1.49$), and even lower d's for face identification tests (challenging subordinate, e.g., "face1" vs "not face1"; mean $d' = 0.50$). Second, human object recognition ability drops as variation in object view (position, scale, pose) increases: mean $d' = 2.39$ for low variation, 1.89 for medium, and 1.50 for high variation. Although these results show that humans are not completely invariant, they confirm that humans tolerate significant amounts of object view variation. Figure 3*a* also shows that tolerance interacts with shape similarity, with humans being the least tolerant for the most difficult subordinate tasks (Biederman and Gerhardstein, 1993; Tarr and Bulthoff, 1998; Tjan and Legge, 1998). We note that these measurements of human object recognition ability were designed and performed independently of any neuronal data collection. Next we wondered which, if any, candidate neuronal linking hypotheses would predict the human pattern of behavior over all 64 tests (Fig. 3*a*).

Specific monkey IT-based linking hypothesis predicts human core object recognition behavior

Before delving into the large space of linking hypotheses that we explored, we start by summarizing our main result. Our analyses revealed that the LaWS of RAD IT linking hypothesis produced a pattern of behavioral performance that accurately predicted the observed pattern of human behavioral performance. Figure 4 shows the predictions of a specific instance of the LaWS of RAD linking hypothesis based on 128 IT neuronal sites, with the re-

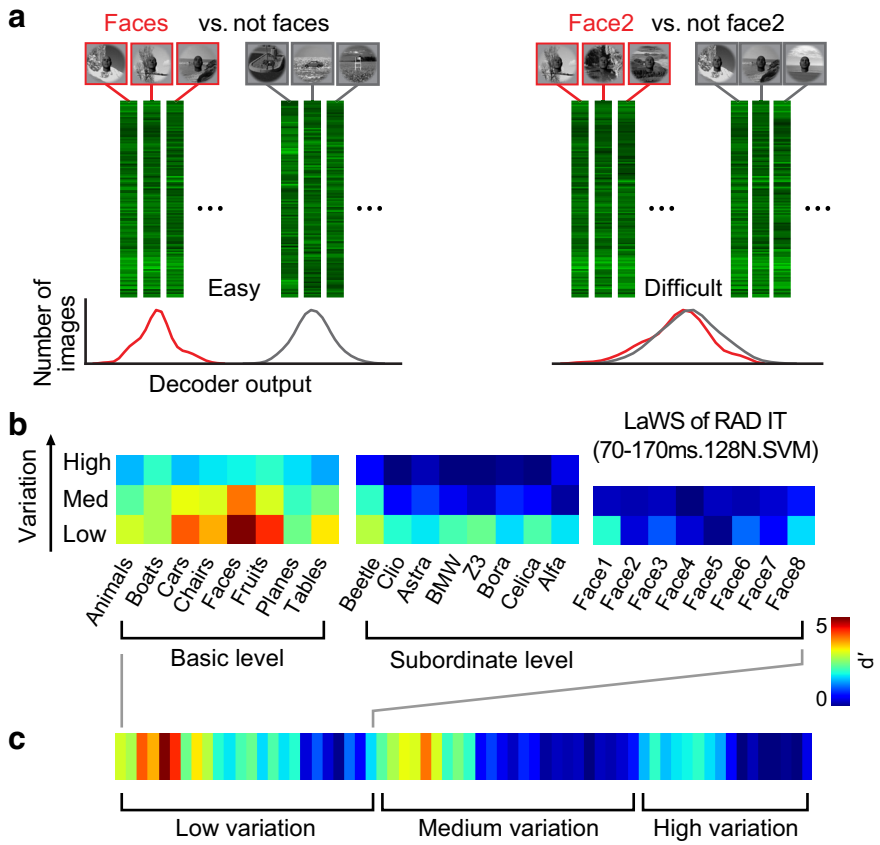

Figure 4. Predicted performance pattern of an example LaWS of RAD IT neuronal linking hypothesis. In this example, the hypothesized neural activity that underlies behavior is as follows: in IT, from 128 units, mean firing rate, in a time window of 70–170 ms; and the decoder is an SVM decoder. ***a***, Based on the aforementioned features of neural activity, a depiction of the outputs of two example decoders for two tasks from two different task sets. For each task set (basic categorization, subordinate identification) and each variation level (low, medium, high), we randomly divided our image responses into "training" and "testing" samples. We used the "training" samples, depicted by the green response vectors, to optimize eight "one-vs-rest" linear decoders. The performance of each decoder was then evaluated on the "testing" images. The red and black distributions summarize the response output of two such decoders to a sample of "testing" images. ***b***, Predicted pattern of behavioral performance for all 64 behavioral tests. To generate these predictions, we constructed an 8-way decoder for each of the three task sets. Analogous to what the human observers were asked to do, for each task set, we applied all eight decoders and scored the decoder with the largest output margin as the behavioral "choice" of the linking hypothesis. Our final d's are the average of at least 50 iterations of randomly picked train/test splits. Similar to Figure 3, the color matrices depict predicted performance (d') for this example linking hypotheses for all task sets and variation levels (64 predicted d' values). ***c***, To facilitate comparison among different linking hypotheses and with human behavior (see Fig. 5), we strung out the color matrices into a color vector grouping task sets at each variation level.

sponse averaged over a 70–170 ms time window after image onset. The pattern of predictions for the 64 recognition tests is statistically indistinguishable from the pattern of human behavior (Fig. 3*a*) and is clearly superior to an identical LaWS of RAD linking hypothesis based on 128 V4 neuronal sites (comparisons quantified in Figs. 5*a*,*b* and 7).

In total, we collected the responses of 168 spatially separate neuronal sites in IT (M1: 58, M2: 110) and 128 sites in V4 (M1: 70, M2: 58). We pooled neuronal sites across IT because we did not see any strong differences between its subdivisions (PIT, CIT, AIT; see Fig. 9). We measured each site's spiking response pattern to each of 5760 images drawn from the same pool used in the human psychophysical testing. Each image was presented at least 28 (typically >47) times per site (i.e., a total of ~250,000 visual stimulus tests at each site). We could not collect this large volume of data in a single day; it required ~30 d of recording in each animal. The initial linking-hypotheses that we explored were based on MUA and assumed stability of that measure at each recording site over the 30 recording days (Chestek et al., 2011). We then investigated how our main finding might change if we used SUA response data instead, and examined our assumption of the stability of each recording site over days (see Materials and Methods and Fig. 8). We also considered the fact that we only sampled a small number of IT neuronal sites (relative to the millions of neurons in IT cortex). Although these factors are important for estimating the number of neurons needed to predict behavior, they turned out to have little impact on our main finding.

Candidate linking hypotheses that might predict object recognition behavior

A candidate linking hypothesis that aims to predict the observed pattern of human recognition accuracy must have at least two components: (1) a specification of the exact "features" of neural activity that are relevant to behavior (i.e., neuronal code) and (2) a specification of a biologically plausible mechanism that translates that neural code to a behavioral choice on each trial.

Based on the existing literature, the "features" of neural activity of high interest include: the tissue region where the neuronal responses are found (V4, IT, IT inside "face patches," IT outside "face patches"), the size of the neuronal population (number of neural sites or units), the temporal window over which the responses are considered, the temporal grain of those measurements (e.g., spike timing codes vs rate codes), and consideration of the so-called "trial-by-trial" population-wide correlation of activity. One can imagine many possible variants and combinations of these ideas, not all of which can be fully explored in a single study, but we aimed to specify and then test some of the most widely believed ideas. That is, we used our data to measure the code specified by each hypothesis and then we asked how well that code predicted the measured object recognition performance.

To compute the predictions of each code, a linking hypothesis must also specify a biologically plausible mechanism (decoder) that translates the measured neural response features into an object label on each trial. For example, a "car" decoder translates the measured neuronal features into the binary decision: is a car present in the image or not? In this study, we tested simple perceptron-like decoders (i.e., linear classifiers, linear discriminants), each of which computes a simple weighted sum of the features in the proposed population code. Although this study is agnostic with respect to how this type of decoder might be implemented in downstream brain areas (e.g., PFC, Freedman et al., 2001; perirhinal cortex, Pagan et al., 2013), it is known to be biologically plausible: each synapse on a hypothetical downstream neuron corresponds to a "weighting" on part of the neu-

ronal code and the neuron's output as determined by the weighted sum of all its inputs corresponds to a decision by the decoder (Shadlen et al., 1996). Different types of decoders correspond to different assumptions about how those downstream neurons learn the synaptic weights (e.g., in humans, this might correspond to learning to map visual inputs to specific words in the lexicon). In this study, because we were primarily interested in the neuronal features that best predict adult object recognition performance, our approach was to start with a simple, well known decoder, hold that idea constant, and then later explore different types of decoders to determine their impact on our conclusions (see Fig. 10). All performance measures reported here are based on neural responses to images that were never previously seen by the decoder (i.e., cross-validation; see Materials and Methods).

In sum, to test each conceptual neuronal linking hypothesis, we: (1) instantiated the idea by measuring the proposed neural code in the population data, (2) learned a hypothesized downstream decoder (e.g., one for each of the 24 noun labels) that takes that neuronal code as input and finds the simple weighed sum that gives the highest performance on that test (see Materials and Methods for details), and (3) computed the behavioral predictions of that hypothesis for each of the 64 tests using previously unseen images.

To the extent that each neuronal linking hypothesis predicts a different pattern of behavioral performance across the 64 tests, not all linking hypotheses can accurately predict the observed pattern of human behavioral performance. A priori, it was also possible that none of the linking hypotheses would accurately predict the human pattern of behavior; for example, we may not have sampled enough neurons to reveal that a hypothesis is sufficient or perhaps monkey and human performance patterns are different and thus no linking hypothesis tested on monkey neuronal codes can predict human patterns of performance. Nonetheless, we reasoned that we could use the strategy of comparing the predicted vs actual object recognition performance of each neuronal linking hypothesis to infer which hypothesis corresponds most closely to the mechanisms at work in the brain.

In total, we tested 944 linking hypotheses, in each case varying the number of neural sites included, thereby translating that conceptual linking hypothesis into an exact specification that makes falsifiable

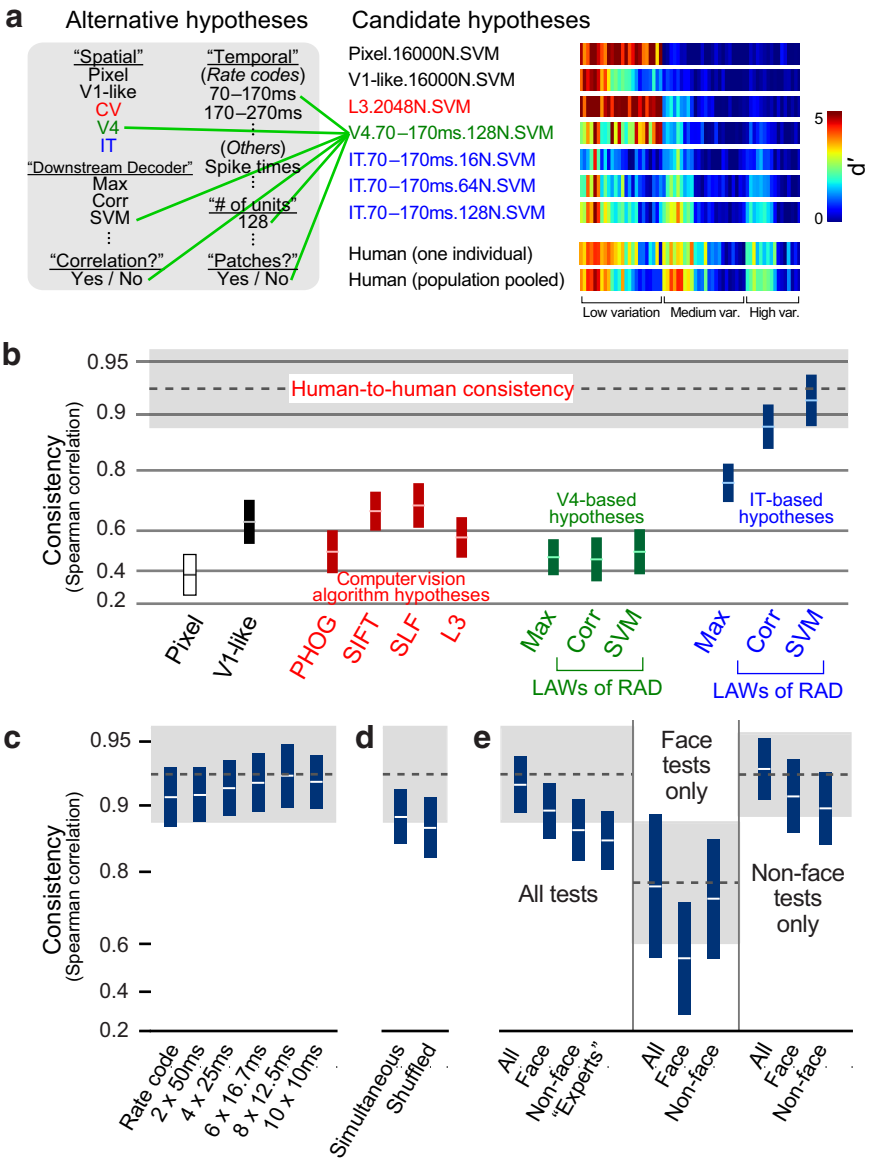

Figure 5. Candidate linking hypotheses. ***a***, Candidate linking hypotheses that we explored were drawn from a space defined by four key parameters: spatial location of sampled neural activity, the temporal window over which the response of our units was computed (mean rate in this window), the number of units, and the type of hypothesized downstream decoder. Each candidate linking hypothesis is a specific combination of these parameters. For example, in green is a V4-based linking hypothesis with a temporal window of 70–170 ms that includes 128 neural sites and uses an SVM decoder. The predicted performance of each linking hypothesis for each behavioral test is depicted as a color vector where blue signifies low predicted performance ($d' = 0$) and red signifies high predicted performance ($d' = 5$). The goodness of each linking hypothesis can be visually evaluated by comparing its color pattern with that of the human population. ***b***, Consistency. To quantify the ability of each linking hypothesis to predict the pattern of human performance (i.e., the similarity between color vectors in ***a***), we computed the Spearman rank correlation coefficient between predicted performance and actual (pooled human, 104 subjects) across all task d's. Median human-to-human correlation is indicated by the dashed line (median Spearman correlation coefficient of 0.929). The gray region signifies the range of human-to-human consistency (68% CI = [0.882, 0.957]). Each bar represents a different candidate linking hypothesis (bar length is proportional to task-induced variability). For pixel features (open symbol), V1-like features (filled black symbol), and computer vision features (red filled symbols), we picked the linking hypothesis that performed best. For neural features (V4 (green) and IT (blue), we matched the number of units at 128. Only bars that enter the gray region correspond to linking hypotheses that successfully predict human behavior. Within the context of IT-based linking hypothesis, we explored finer grain temporal codes (***c***). We also took advantage of our simultaneous multielectrode array recording to assess the impact of trial-by-trial firing rate correlation on the pattern of performance predicted by our most successful linking hypothesis (***d***). We considered the idea of a modular IT linking hypothesis with different subregions of IT being devoted exclusively to certain kinds of tasks (***e***). First, we compared the performance of "face patch" sites to "nonface patch" sites on all tests. We then stitched together an "expert linking hypothesis" in which each test is performed by neuronal sites that are tailored to that test (e.g., "face" detection is only done by "face neurons" whereas "car" detection is done by nonface neurons). To be complete, we compared the performance of our different modular IT linking hypotheses on both face tests only ($n = 17$ of the 64 tests) and nonface tests only. As in ***b***, pattern of performance was always compared with human-to-human consistency indicated by the gray region.

predictions. For example, one specific linking hypothesis that we tested was as follows: learn simple weighted sums of the mean firing rates across 128 IT neural sites, distributed across IT measured in a 70–170 ms time window, ignore trial-by-trial correlations (Fig. 4). To facilitate visual inspection, the behavioral predictions of this linking hypothesis are strung out in a single color-coded vector (Fig. 4*c*). Most candidate linking hypotheses produced different predicted patterns of behavioral performance; sometimes, these differences were small, but they were often dramatic (see Fig. 5*a* for examples). For visual comparison, Figure 5*a* also shows human performance on the same 64 object recognition tests (data from Fig. 3*a*, replotted), illustrating that some candidate linking hypothesis lead to very poor predictions of the pattern of human performance, whereas others lead to surprisingly good predictions.

Quantifying the goodness of a linking hypothesis: consistency

Following previous work (Hyvarinen et al., 1968; Newsome et al., 1989; Connor et al., 1990), we reasoned that the neuronal linking hypotheses that are the most likely to correspond to the mechanisms at work in the brain are those that produce the most quantitatively consistent relationship with the human behavior (i.e., the linking hypothesis's pattern of colors in Fig. 5*a* should best match the pattern of colors in Fig. 3*a*). To quantify that consistency, we computed Spearman's rank correlation coefficient over the 64 d's (Yoshioka et al., 2001).

The most stringent application of this method is that, for a linking hypothesis to remain viable, it must produce behavior that is indistinguishable from the behavior of individual subjects. Based on this stringent criterion, all V4-based linking hypotheses that we tested failed to accurately predict the observed human behavioral pattern (Figs. 5*b*, 6), as did all V1-based linking hypotheses. For comparison, Figure 5, *a* and *b*, also shows the predictions of linking hypotheses based on populations of baseline computer vision features, all of which failed to predict the pattern of behavior (see Materials and Methods for details). Despite our best efforts, we found that the V4- and V1-based linking hypotheses could not be "rescued" by increasing the number of neurons in the linking hypothesis or by changing the type of decoder (i.e., learner; Figs. 5, 6). We also considered the possibility that V4-based linking hypotheses might have been handicapped by receptive field limitations of our neural sampling. In particular, we narrowed our images to only those with objects presented in the contralateral field or at the center of gaze. Although V4 populations showed the expected pattern of higher d's for contralaterally presented objects, neither test substantially improved the ability of V4-based linking hypotheses to predict the pattern of human behavior (see Materials and Methods) even though these same V4 populations often outperformed humans in some of the behavioral tests (see below). These results do not suggest that V1 and V4 play no role in object recognition behavior, but rather that neural representations (i.e., codes) conveyed by those areas do not directly underlie object recognition behavior, which is compatible with previous suggestions (Sheinberg and Logothetis, 1997; Brincat and Connor, 2004; Rust and DiCarlo, 2010; DiCarlo et al., 2012). These results also show that the approach we used, the combination of images and tasks, is a powerful test of neuronal linking hypotheses that cannot easily be "passed" by lower-level (e.g., V1) or even mid-level (V4) representations.

In contrast to the results in V4 and (simulated) V1, we found that some IT-based linking hypotheses accurately predicted the behavioral pattern of human observers. For example, based on previous work (Hung et al., 2005), one of the first specific linking

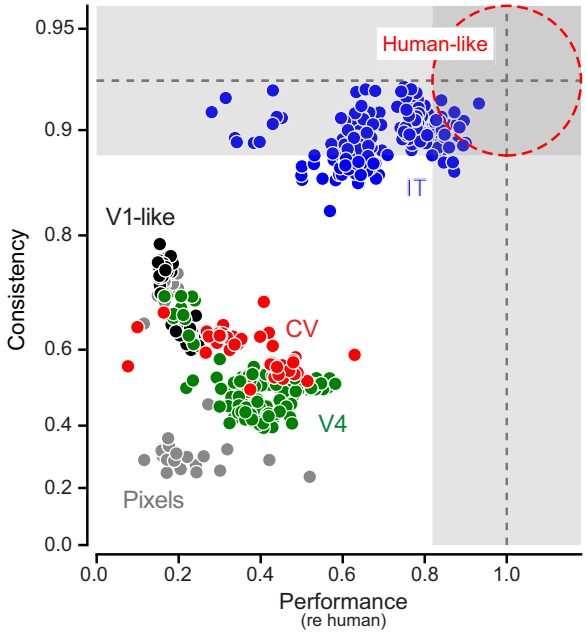

Figure 6. Exploring a large set of linking hypotheses. The *y*-axis shows consistency (defined in Fig. 5*b*) and the *x*-axis shows performance—the median of the ratio between predicted and actual (human) d' across all 64 tests. In total, we tested 944 types of linking hypotheses, varying the number of neurons/features in each case, for a grand total of 50,685 instantiations considered. Here, we show the results of 755 of those hypotheses. The result of each specific instantiation is shown as a point in the plot with color used to indicate the "spatial" location of the features (IT, V4, V1, or computer vision). We show these examples to illustrate the parameters that we varied, which included spatial location, temporal window, number of units, type of decoder, and a variety of training procedures and train/test splits (see Fig. 10*a*). The horizontal dashed line indicates the average human-to-human consistency and the horizontal gray band represents variability in human-to-human consistency. The vertical dashed line indicates the average relative human-to-human performance and by definition is at 1 and the vertical gray band shows the human-to-human variability in relative performance. Any linking hypothesis that falls in the red dashed circle is perfectly predicting human performance on these 64 tests. Note that much of the scatter in the IT-based linking hypotheses (blue) is due to varying the number of neural sites, as illustrated in Figure 7*b*.

hypotheses that we tested was as follows: the mean firing rate of each IT neurons in a 70–170 ms time window, where IT neurons are sampled in a distributed manner over IT cortex (i.e., ignoring IT spatial substructure such as "face-patches") and ignoring correlations across the population (Fig. 5*d*). We tested this linking hypothesis using different numbers of IT neural sites and were surprised to find that, once we included ~100 sites, this IT-based linking hypothesis was not only a more accurate predictor of human behavior than other hypotheses (e.g., V4-based linking hypotheses), but its predictions were statistically indistinguishable from the measured pattern of human d's (linking hypotheses that pass into the gray region in Fig. 5*b*). Following up on this result, we also found this simple IT-based linking hypothesis continued to accurately predict the pattern of human object recognition ability even when we varied the number of neuronal sites participating in the linking hypotheses (>64 sites), the type of decoder used, and the training provided to the decoder (Figs. 7, 8, 9, 10).

We explored several other IT-based linking hypotheses that have been suggested in the literature. First, we considered the idea that trial-by-trial correlations in neuronal firing across the IT population might be important to consider when asking if a neuronal linking hypothesis is consistent with behavior (Zohary et al., 1994; Averbeck et al., 2006; Liu et al., 2013). Because we had

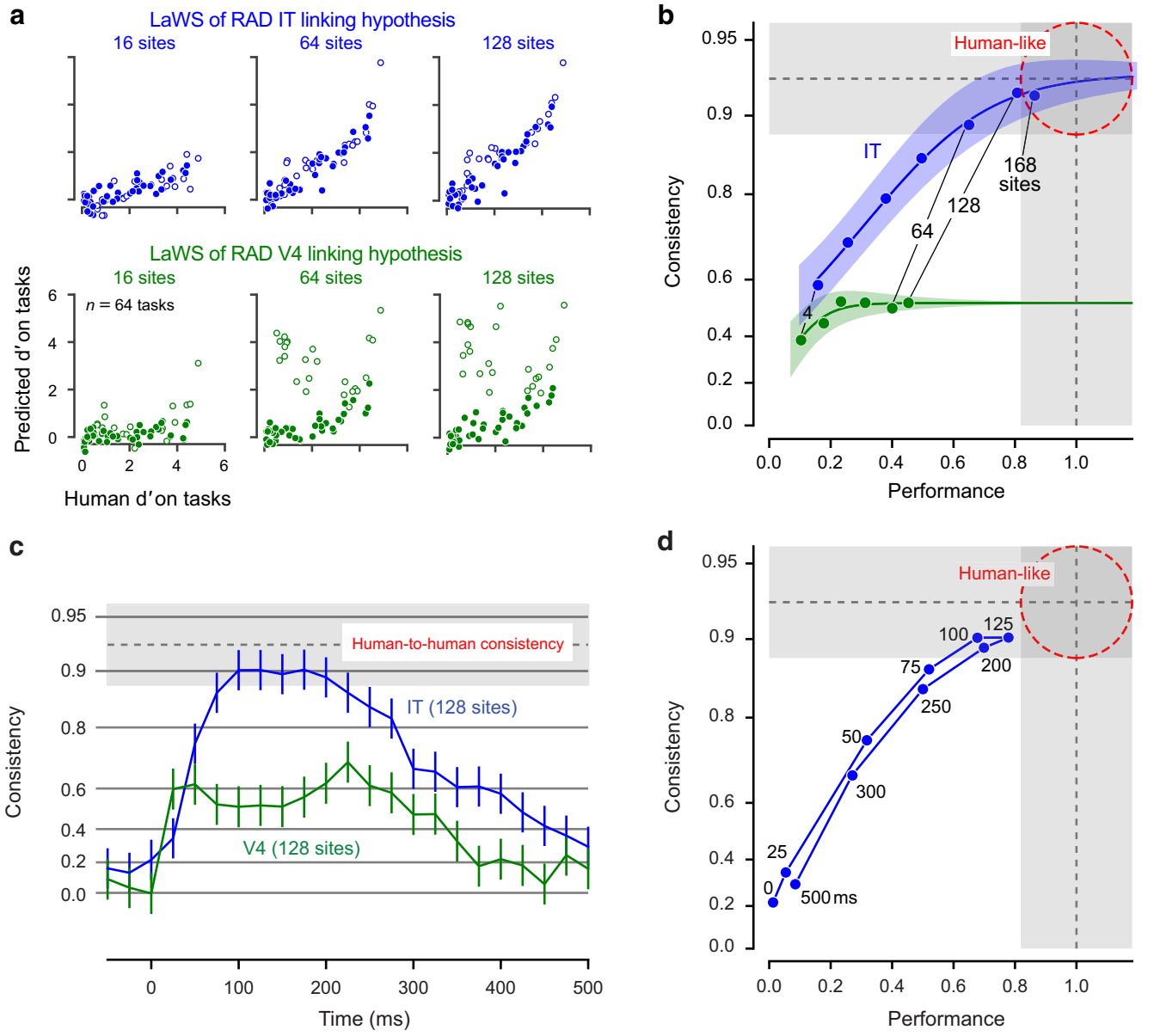

Figure 7. Effect of number of units and temporal window on consistency and performance. Here, we show the results for the LaWS of RAD linking hypotheses (see text), but results are qualitatively similar for other hypotheses. ***a***, Scattergrams show predicted performance (d') for two neuronal linking hypotheses, IT (blue) and V4 (green), plotted against the actual human performance on all 64 tests: low variation (open circles), medium and high variation (filled circles). The number of units increases from 16 neural sites (left) to 128 neural sites (right). For each linking hypothesis, we also computed its performance: the median of the ratio between predicted and actual human performance across all d's for all 64 tests. ***b***, Performance (defined in Fig. 6) versus consistency for the V4- and IT-based linking hypotheses as a function of the number of (trial-averaged) units. The curve fits are [IT, $r^2 = 0.996$; V4, $r^2 = 0.91$] and they predict that ~529 IT trial-averaged neural sites and ~22,096 V4 trial-averaged neural sites would match human performance under the LaWS of RAD linking hypothesis. ***c***, Consistency for different temporal windows of reading the neural activity. Each point is computed with a 100-ms-wide window and the *x*-axis shows the center of that window. The number of trial-averaged neural sites was fixed at 128. ***d***, Consistency versus performance for the LaWS of RAD IT linking hypothesis at several progressive temporal windows with the center location starting at the time of image onset (0 ms) and up to 500 ms after image onset. The width of the temporal window was fixed at 100 ms (code details are same as ***b***, except the number of trial-averaged neural sites was fixed at 128).

collected responses at many of our neuronal sites simultaneously, we were able to compare neuronal codes produced across the population on actual single trials, with codes produced on artificial single trials in which any population correlation structure is removed by shuffling the trials (e.g., so the responses of IT unit 1 on presentation p of image i are considered along with the responses of IT unit 2 on presentation q of image i). We found that a LaWS of RAD IT linking hypothesis that maintained the trial-by-trial population correlation structure had no increased (or decreased) ability to explain the pattern of human behavior, even when lowering the number of neurons so that we might be able to see that increase (Figs. 7, 8).

Second, we considered the idea of finer-grained temporal codes. To do this, we took the simple 70–170 ms poststimulus time window (above) in which the LaWS of RAD IT linking hypothesis was highly predictive and broke it into successively smaller and smaller time windows, giving each learned decoder full access to the neural response in all such time windows. Because all of the same spiking information in these finer-grained temporal codes is still available to each decoder, this approach can only maintain or increase performance on each of the 64 behavioral tests (until data limits are reached). However, because accuracy on some tests might improve relative to others, it could increase, decrease, or have no effect on the consistency

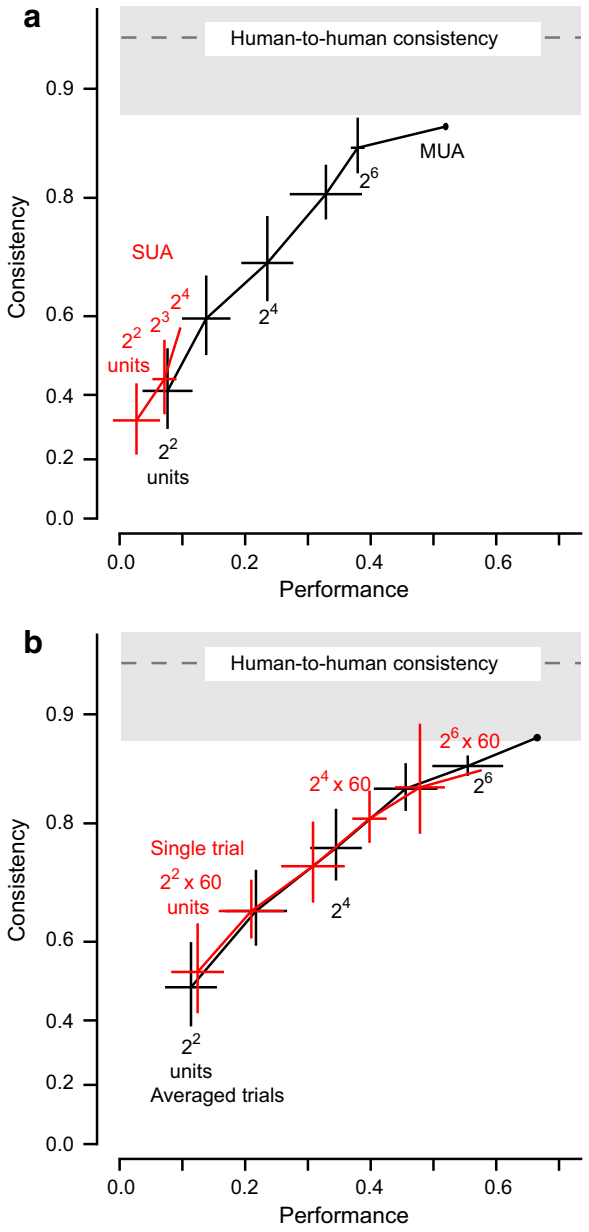

Figure 8. ***a***, SUA versus MUA linking hypotheses. We used a profile-based spike sorting procedure (Quiroga et al., 2004) and an affinity propagation clustering algorithm (Frey and Dueck, 2007) to isolate the responses of 16 single units from our sample of 168 IT neuronal sites. The minimum signal-to-noise ratio (SNR) for each single unit cluster was set to 3.5, with SNR defined as the amplitude of the mean spike profile divided by root mean square error (RMSE) across time points. Consistency with the human pattern of performance versus performance for SUA (red) and MUA (black). We estimate that twice as many neurons are needed so that the consistency-performance relationship of our SUA linking hypothesis matches that of our MUA linking hypothesis. All parameters and training procedures of SUA- and MUA-based linking hypotheses were identical (performance was based on the average of five repetitions using a CC in which the units were randomly divided into nonoverlapping groups to estimate error from independent sampling of units). ***b***, Single trial versus averaged trials linking hypotheses. Because human subjects were asked to make judgements on single image presentations, we also explored a "single trial" training and testing analysis in which we treated the responses of the neural units to each images presentation as a new and independent set of neural units (i.e., "unrolled" the trial dimension into the unit dimension). Consistency versus performance for the single-trial (red) and the averaged-trial (black) LaWS of RAD linking hypotheses (based on a correlation decoder). We estimate that ~60 times as many neurons are needed so that the consistency-performance relationship of our single-trial linking hypothesis matches that of our averaged-trials linking hypothesis. Error bars are SDs induced by independent sampling of units as in ***a***.

of the pattern of performance over all 64 human behavioral tests. The results showed that, relative to the simple 100 ms window mean firing rates in the LaWS of RAD linking hypothesis, these more complex, finer-grained IT temporal codes led to no measurable change in consistency with the pattern of human behavior.

Third, we considered modular IT linking hypotheses in which different subregions of IT are devoted exclusively to certain kinds of tasks. The strongest experimentally motivated example of a modular linking hypothesis is that certain spatial regions of IT (face "patches" in monkeys; fusiform face area, occiptal face area in humans) are devoted to certain types of "face-related" tasks, such as face discrimination (one face vs others) and face detection (faces vs other categories). Our data allowed us to examine such hypotheses because 19 of our 64 tests are face-related tasks and we could label ~19% of our IT neural sites as likely belonging to one or more of the 6–10 IT face patches (based on the high purity of these regions for units that have high face vs nonface object selectivity; Tsao et al., 2006). We first note that our findings are consistent with weaker forms of modularity of face processing, such as spatial clustering of neural sites that are most important for face detection. Indeed, it was not surprising (as it is nearly by definition) that, of the IT sites that were weighted most strongly by the decoders (i.e., the top 5% most heavily weighted) in our three face detection tests, 87.5% of those were face-patch-likely sites. More interestingly though, we also found that only 12.5% of the most highly weighted sites in our 16 face discrimination tests were face-patch-likely sites, arguing that face discrimination might not rely exclusively on IT face-patch tissue. To test a stronger form of the face modularity hypothesis using the consistency approach of this study, we investigated whether neuronal linking hypotheses based only on the face-patch-likely population of sites were more consistent with the pattern of human performance on face-related behavioral tests (compared with linking hypotheses based on all of IT or based only on face-patch-unlikely populations within IT; Fig. 5*e*). We found that this did not significantly change the accuracy of the behavioral fits; if anything, the trend suggested a decreased accuracy. In sum, whereas our results are consistent with weaker forms of the face-modularity linking hypothesis [i.e., face detection tasks are best performed by "face (detection) neurons" that are spatially clustered; Tsao et al., 2006; Issa and DiCarlo, 2012], our data find no support for the stronger form of the face-modularity hypothesis (i.e., all face-related tasks exclusively depend on the responses of neurons in face patches). However, our data do not falsify that strong form either.

Goodness of a candidate linking hypothesis: performance

Although the consistency metric evaluates the similarity between the pattern of d's predicted by each candidate linking hypothesis and the measured human d's, we next asked: what number of neurons is required for a LaWS of RAD IT linking hypothesis to account for the actual d's across all our 64 tests? In particular, one can imagine neuronal linking hypotheses that are highly predictive of the pattern of d's over the 64 tests (as in Fig. 5), but with absolute levels that are far below the measured human d's (see Fig. 1*b* for a schematic demonstration of correlated but unequal d's). Indeed, we found examples of such linking hypotheses (see blue points in Fig. 6 that are within the top gray band but outside of the red dashed circle). We found that, for both V4 and IT-based codes, once the number of neural sites was greater than ~100, measures of consistency were quite robust to further increases in the number of neural sites in the code. However, performance, the median of the ratio between predicted and actual

(human) d' across all 64 tests, of any specific neural code was strongly dependent on the number of neuronal sites. For example, whereas we found it effectively impossible to vary the number of neural sites to make (for example) a V4-based linking hypothesis match the human pattern of performance (Fig. 7*b*), for many V4-based linking hypothesis, we could, by extrapolation, estimate the number of neurons that could, in principle, match the median human d' over the 64 tests.

Effect of number of units on consistency and performance

We systematically explored the effect of changing the number of neural sites on consistency and performance. This is illustrated in Figure 7 for two families of linking hypotheses—the simple LaWS of RAD IT linking hypothesis family reviewed above, and the simple LaWS of RAD V4 linking hypothesis family. For both families, median predicted performance increased as the number of sites increased; however, only the LaWS of RAD IT linking hypothesis became fully consistent with human performance. That is, with 128 neuronal sites (or more), the LaWS of RAD IT linking hypothesis shown in Figure 7 perfectly predicted the entire pattern of performance over all 64 tests in that the Spearman correlation (Fig. 7*a*,*b*) was indistinguishable from the human-to-human consistency (the horizontal dotted line in Fig. 7*b*; the gray region indicates the variability of individual human subjects).

Figure 7*a* also illustrates why the non-IT-based linking hypotheses that we tested failed to explain the pattern of human performance. In particular, it shows that the LaWS of RAD V4 linking hypothesis fails both because it cannot achieve high d's on some tests (e.g., high variation tests, green filled circles in Fig. 7*a*) and because it achieves d's that are better than humans in other tests (e.g., some low variation tests, green open circles in Fig. 7*a*). Increasing the number of neurons participating in the LaWS of RAD V4 linking hypothesis cannot fix this obvious discrepancy with behavior and the result argues against the idea that we did not collect sufficient information from V4 neurons. In sum, distributed, learned V4 population rate codes do better than humans on some particular behavioral tests, but they fail to produce the human pattern of d's over all 64 tests.

Sufficient single-trial, single-unit population linking hypotheses

As shown in Figure 7, both the consistency and performance of IT-based linking hypotheses are dependent on the number of neural sites assumed to be participating in the behavior. The plot shows that only a small number (hundreds) of neural sites are needed before the consistency of LAWs of RAD IT hypothesis plateaus; that is, that linking hypothesis achieves a pattern of performance that is indistinguishable from humans after it includes ~100 sites (*y*-axis in Fig. 7*b*) and the inclusion of more neural sites does not improve consistency. However, another constraint on the number of neuronal sites comes from how the mean performance of a specific linking hypothesis compares with the mean performance of human subjects (*x*-axis in Fig. 7*b*; see the caption for the definition of performance). Unlike consistency, performance is an unbounded metric that depends on signal-to-noise ratio. Too few neuronal sites lead to median predicted performance that is below observed performance and too many lead to performance that is superior to behavior. This offers the opportunity to find the number of neural sites where the linking hypothesis matched human performance (Fig. 1*b*, upper right). However, to estimate that number of neurons, it becomes very important to consider exactly how the hypothesis is implemented and its relationship to brain circuitry. In particular, we and others assume that neurons in downstream brain areas can listen to the spikes of some number of single neurons (e.g., neurons in IT) and produce, on each behavioral trial, a guess as to the object label (the test we asked the humans to perform; Fig. 3*a*). The neural data and analyses used to generate Figures 4, 5, 6, and 7 differed from this assumption in two ways: (1) we did not distinguish single units from multiunits and (2) we averaged the responses of each neural site over many repetitions (typically 50, minimum 28). Neither of these details substantially altered our conclusions about the behavioral consistency of LaWS of RAD IT linking hypotheses. However, they are important for estimating how many single neurons would be needed to match human-level accuracy on single image presentations.

Figure 8*a* examines the difference between MUA and the activity of sorted SUA. Linking hypotheses based on SUA and MUA IT data shared highly comparable consistency-performance relationship, except that SUA linking hypotheses required approximately twice as many neural sites to reach a similar level of consistency or performance. This similarity is perhaps surprising (see Discussion), but is compatible with previous work that examined the same issue (Hung et al., 2005).

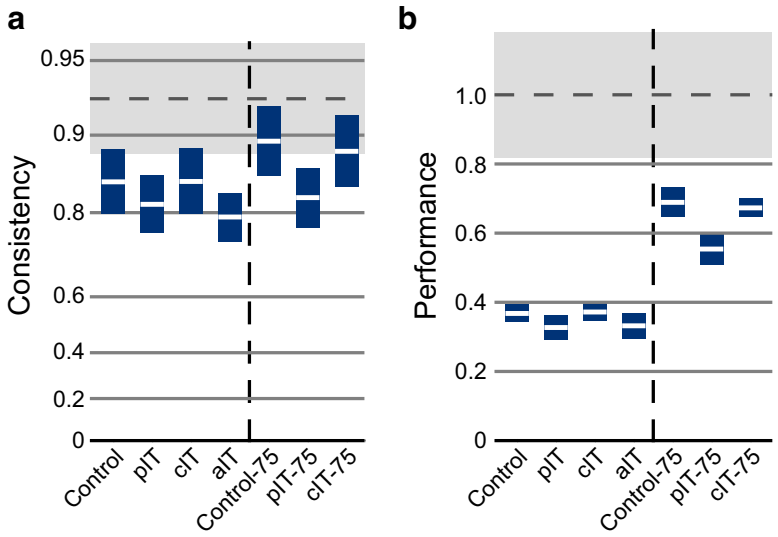

Figure 9. No significant difference in consistency and performance of IT subpopulations was seen when parsed based on anatomical subdivision: PIT versus CIT versus AIT. Based on anatomical landmarks, we could conservatively divide our population of 168 IT neural sites into the following: 76 in PIT, 75 in CIT, and 17 in AIT. ***a***, Comparison of the consistency values for IT populations when neural sites respected anatomical boundaries (PIT vs CIT vs AIT) in contrast to a "control" populations in which the sites were randomly picked from all three anatomical subdivisions. There was no significant difference between the IT populations regardless of whether we restricted our population to 17 neural sites (limiting our analysis to the number of neural sites in AIT our least sampled anatomical subdivision) or expanded to 75 neural sites and compared PIT and CIT. Similarly, performance (***b***) showed no significant differences between the different IT populations. It is important to note that the decrease in consistency and neural performance is expected based on the smaller population sizes (see Fig. 7*b*). Consistency and performance were computed based on our typical 70–170 temporal window using an SVM decoder.

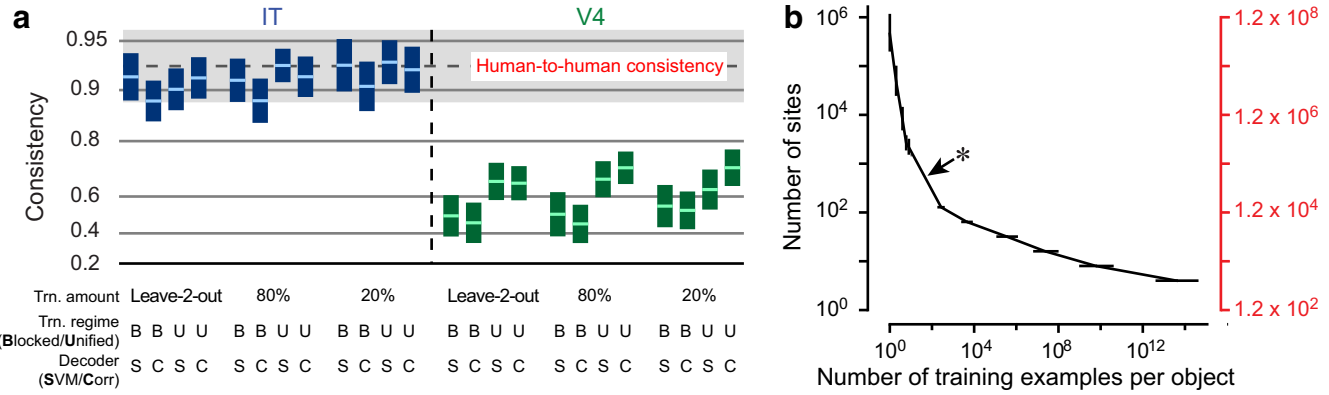

Figure 10. ***a***, Effect of the training procedure. Shown are consistency values for LaWS of RAD V4 and IT linking hypotheses under different training procedures. The number of units was fixed to 128 units and the temporal window was 70–170 ms after the onset of the image presentation. Two types of decoders were tested (SVMs and CCs). We also varied the number of images used to train the decoder (leave-2-out: for each class, all images but two were used as the training set and the remaining two were used for testing; 80%: 80% of images were used for training, and the held-out 20% were used for testing; 20%: similar to 80%, but 20% were used for training and 80% for testing). In the blocked training regime, the training and testing of a decoder was done for each variation level separately. For the unified training regime, the decoders were trained across all variations and tested on each variation level separately. ***b***, Trade-off between the sufficient number of units and the number of training images per object for the LaWS of RAD IT linking hypothesis (in which the temporal window was fixed at 70–170 ms and SVM decoders were used). In each data point, the performance of the linking hypothesis was projected to reach the human-to-human consistency (within the subject-to-subject variability) and the human absolute performance (relative performance of one). On the *y*-axis, the numbers shown in black indicate the projected number of repetition-averaged, multiunit neural sites that are sufficient and the numbers in red indicate the number of single-trial, single-unit sites that are sufficient (120× larger). For example, the asterisk indicates a LaWS of RAD IT linking hypothesis of ~60,000 single units discussed above and the plot shows that it would require ~40 training examples per object to learn *de novo* (with a 68% CI of ~[30, 60]; data not shown in the plot).

Figure 8*b* explores the issue of averaging and compares the results of a simple model of single trial decoding to the results of decoding while averaging across all available trials. Although we did not expect this analysis choice to change our conclusions about the behavioral consistency of LaWS of RAD IT linking hypotheses, we expected that it would affect the estimated number of IT neurons that must participate in that linking hypothesis to achieve human-level performance on single trials (because averaging improves the signal-to-noise ratio of each neuronal site). The single-trial analysis in Figure 8*b* gives a consistency-performance relationship similar to that of averaged-trial analysis if ~60 times as many IT units are provided. That is, we estimate that ~60 independent IT neural sites (operating in parallel) are sufficient to stand in for a single, "repetition-averaged" neural site and this estimate accounts for how neuronal variability ("noise") affects both the decoding (e.g., as in Shadlen et al., 1996) and the learning of the decoder (see more below).

Together, the analyses presented in Figure 8 converge to suggest that spike counts from ~60,000 (529 repetition-averaged IT multiunits × ~2 × ~60; Fig. 7*b*) distributed single units in IT cortex can, when read with simple, biologically plausible downstream neural decoders, perfectly predict both the behavioral pattern of performance and the median level of performance over all 64 tested object recognition tests. This number is an extrapolation because our methods are not yet capable of recording that many IT neurons and other factors such as "noise correlation" might alter that estimate (see Discussion). Furthermore, because performance depends on parameters of how the code is learned to be read (decoded), this estimate could be somewhat higher or lower, as analyzed in detail in Figure 10. However, we note that this number is far less than the total number of neurons estimated to project from IT to downstream targets (~10 million; DiCarlo et al., 2012).

Effect of time window on consistency

To further explore the precise parameters of the LaWS of RAD IT family of linking hypotheses, we varied the starting time and duration of the time window over which the mean rate was read from the IT population (Fig. 7*c*,*d*). We found that the LaWS of RAD IT linking hypothesis begins to be highly consistent with behavior at a center latency of 100 ms (time window of [50, 150] ms after image onset) and that consistency remains at a high plateau for nearly 100 ms before dropping off. During this entire plateau, the predicted pattern of performance of this linking hypothesis is statistically indistinguishable from the human pattern of performance. For comparison, all LaWS of RAD V4 linking hypotheses that we tested failed to pass this consistency test for all temporal windows.

Discussion

We propose a framework for comparing neural responses with behavior. Instead of qualitatively comparing performance on a selected set of conceptual tasks, we devised a "Turing" test—a battery of behavioral tests that explore the range of human subjects' capabilities in core object recognition. This operational definition of object recognition provided a strong consistency test by which we could quantitatively evaluate different neuronal linking hypotheses that might explain behavior. As expected, many neural (and non-neural) linking hypotheses failed to predict object recognition behavior, including: pixel-based codes, V1-like-based codes, multiple computer vision codes, V4-based codes, and several IT-based codes. However, we were surprised to find that the LaWS of RAD IT linking hypothesis family perfectly predicted the human pattern of behavioral performance across all 64 recognition tasks. More precisely, the data argue that a simple rate code (100 ms time scale, [70, 170] ms onset latency) read out on single trials learned from a distributed population of ~60,000 single IT units can fully explain both the pattern and the magnitude of human performance over a large battery of recognition tests.

Initially, we were surprised that this simple linking hypothesis was perfect at predicting the pattern of performance. Nevertheless, we explored other ideas motivated from theoretical considerations (Averbeck et al., 2006) and neuronal response findings (Sugase et al., 1999; Tsao et al., 2006). First, we found that the LaWS of RAD linking hypothesis was not strongly affected by

trial-by-trial correlational structure in the population responses (Fig. 5*d*). We suspect that this is due to the dimensionality of our neuronal populations (>100), combined with the fact that correlational "noise" structure can either increase and decrease performance depending on the layout of the task-relevant "signal" structure in the population representation (Averbeck et al., 2006). Second, we explored finer-grained temporal codes (Fig. 5*c*), which revealed no change in the accuracy of the behavioral predictions. We are careful to note that our results do not imply that trial-by-trial correlational structure is not a limiting factor for some tasks (Mitchell et al., 2009; Cohen and Maunsell, 2010), or that finer-grained temporal neuronal codes in IT are falsified by our data. Instead, our results argue that such ideas do not yet add any measurable value for the real-world-motivated set of object tests explored here.

Our study was not aimed at improving upon the previously documented spatial-clustering of "face neurons" in IT (Desimone et al., 1984; Tsao et al., 2006; Issa and DiCarlo, 2012). However, we did explore the idea that IT is not best considered as a distributed neural representation, but that it consists of at least two spatially segregated parts—"face patches" that are a priori devoted to "face" tasks (part A) and other parts that are devoted to nonface tasks (part B). Our results are entirely consistent with the hypothesis that "part A" neurons are heavily weighted in adult face detection tasks. That is, before learning face detection, downstream neurons accept inputs that are distributed over all of IT, but in the adult, learned state, those downstream readers will most heavily weight neurons that are best at supporting face detection. This hypothesis is consistent with the idea (Tsao et al., 2006) that "face neurons" (and "face patches") are heavily causal in adult face detection behavior (S. R. Afraz et al., 2006; A. Afraz et al., 2015). We also considered a stronger form of domain-specific face processing: that all face related tasks causally depend only on neurons in part A, whereas all other tasks causally depend only on neurons in part B (Tsao et al., 2006). We tested this idea by restricting the parts of IT the downstream decoders are allowed to read from—decoders for face-related tasks can read only from neurons in part A and decoders for all other tasks can read only from part B. Our results showed that such parcelling did not improve the accuracy of the behavioral predictions. Instead, the (nonsignificant) trend was in the wrong direction (Fig. 5*e*). Therefore, our results do not support the strong face modularity hypothesis, but they do not falsify that idea either.

We are not the first to compare neural responses with object recognition behavior. Using shape similarity judgements, some studies have shown agreement between neural representation in monkey IT and perceptual "distances" between parametrized shapes in both monkeys and humans (Op de Beeck et al., 2001; Kriegeskorte et al., 2008). Although pioneering, there is a limit to such qualitative comparisons. Primarily, there is a question as to whether shape similarity is a good surrogate for recognition behavior. However, even if that assumption were granted, previous work did not attempt to rule out V4 or even V1 as viable candidates, nor did it attempt to distinguish among the large space of IT-based linking hypotheses.

Other studies focused on documenting IT's computational prowess at invariant object recognition (Hung et al., 2005; Rust and DiCarlo, 2012). Absolute accuracy was used as the metric, with IT neural populations having a clear advantage over pixels (Hung et al., 2005) and over V4 (Rust and DiCarlo, 2010) in discriminating between objects across limited changes in view. Here, we show that V4-based rate codes are unlikely to directly underlie all object recognition tasks because they outperform humans on some tests and underperform on others. This highlights the fragility of using performance on a single task as a metric for determining which neuronal linking hypothesis underlies behavior. Absolute performance strongly depends on parameters that control the noisiness of a neuronal population (e.g., number of neurons), making it very difficult to expose the key factors of interest (i.e., which neurons and which features of the neuronal response). For example, we here replicate previous work (Zohary et al., 1994; Hung et al., 2005; Cohen and Maunsell, 2009; Rust and DiCarlo, 2010) showing that increasing the number of neurons improves performance on our recognition tests, but we now show that it keeps the relationships between easy and difficult tests the same. Therefore, the pattern of performance across many tests emerges as a more robust measuring stick by which we can evaluate different neuronal codes (Johnson et al., 2002).

Our comparison of nonhuman and human primates deviates from approaches that combine neural recording with behavioral testing in the same subjects (Britten et al., 1996; Luna et al., 2005; Cohen and Maunsell, 2011). It was motivated by our desire to get both high-fidelity behavioral and neuronal population data, a fruitful first-line strategy when a perceptual domain is poorly understood (Mountcastle et al., 1969; Johnson et al., 2002). Such comprehensive characterization of object recognition ability is difficult and time consuming in nonhuman primates and current human fMRI lacks the appropriate spatial and temporal resolution necessary for characterizing neuronal population at the level that we accomplished here (but see Kay et al., 2008; Naselaris et al., 2009).

The fact that monkey neuronal population responses can accurately predict human performance patterns adds evidence to the assumption of highly conserved visual capabilities across the two species (Merigan, 1976; Sigala et al., 2002; Rajalingham et al., 2015). Furthermore, our results show that simple LaWS of RAD in nonhuman primate IT are sufficient to account for human performance, even on object categories outside of the realm of typical monkey experience (e.g., planes, cars, boats, etc.). We interpret this to mean that primates share a generic neural representation of "shape" (Kriegeskorte et al., 2008; Zoccolan et al., 2009) that is suitable for dealing with the difficulties of identity preserving image transformations without being restricted to object categories and a lexicon that is shaped by each subject's real world experience (Freedman et al., 2001). Specifically, our results argue that primates share a nonsemantic IT visual "feature" representation upon which semantic understanding can be learned, which constitutes a performance bottleneck in primate object recognition. This inference is agnostic as to how much of this feature representation is innate versus learned during the statistically shared postnatal experience of primates (Li and DiCarlo, 2008).

Our results set the stage for new directions in linking neurons to object behavior. One natural extension is to obtain more precise behavioral data for the images that we already tested neurally to look closely at the ability of the LaWS of RAD IT linking hypothesis to predict the image-by-image confusion patterns in humans. Another obvious direction is to increase the scope of our images and tasks and explore the effects of crowding, clutter, occlusion, and correlated backgrounds. Both directions will facilitate more stringent neuronal-to-behavioral comparisons and increase the resolution at which neuronal linking hypotheses can be distinguished. Eventually, more comprehensive behavioral tests might force us to turn to more complex underlying neural codes that were not necessary here (e.g., fine-timing or synchrony based

codes; Engel et al., 2001) and might open the door for investigating a role for feedback in tasks that require inference (Kersten et al., 2004; Oliva and Torralba, 2006).

More comprehensive behavioral assays will necessitate conducting them in both humans and nonhuman primates to determine when the cross-species assumption breaks down. As in other sensory areas (Connor et al., 1990; Shadlen et al., 1996; Cohen and Maunsell, 2010), simultaneous recording from behaving animals will reveal a better estimate of the neuronal population size needed for object recognition and produce accurate trial-by-trial performance predictions. The LaWS of RAD IT linking hypothesis reported here brings us a step closer to predicting the impact of direct manipulation of IT neurons on object recognition behavior. In such a framework, future investigations of the behavioral changes in recognition induced by artificial neuronal manipulation (S. R. Afraz et al., 2006; Verhoef et al., 2012; A. Afraz et al., 2015) can be used to further refine IT-based linking hypotheses.

This study sidesteps the important question of how IT neuronal responses are produced. Ongoing work is systematically characterizing the nonlinear transformations from retina through V1, V2, and V4 (Pasupathy and Connor, 1999; Hegdé and Van Essen, 2000; Rust and DiCarlo, 2012; Freeman et al., 2013; Yamins et al., 2014). Those approaches need to be combined with the framework presented here to achieve an end-to-end understanding of the neuronal mechanisms that support core object recognition behavior (DiCarlo et al., 2012).

References

Afraz SR, Kiani R, Esteky H (2006) Microstimulation of inferotemporal cortex influences face categorization. Nature 442:692–695. CrossRef Medline

Afraz A, Boyden ES, DiCarlo JJ (2015) Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. Proc Natl Acad Sci U S A 112:6730–6735. CrossRef Medline

Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nat Rev Neurosci 7:358–366. CrossRef Medline

Biederman I, Gerhardstein PC (1993) Recognizing depth-rotated objects: evidence and conditions for three-dimensional viewpoint invariance. J Exp Psychol Hum Percept Perform 19:1162–1182. CrossRef Medline

Biederman I, Gerhardstein PC, Cooper EE, Nelson CA (1997) High level object recognition without an anterior inferior temporal lobe. Neuropsychologia 35:271–287. CrossRef Medline

Brincat SL, Connor CE (2004) Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat Neurosci 7:880–886. CrossRef Medline

Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA (1996) A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis Neurosci 13:87–100. CrossRef Medline

Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, Majaj NJ, DiCarlo JJ (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol 10:e1003963. CrossRef Medline

Chang C-C, Lin C-J (2011) LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol 2:1–27.

Chestek CA, Gilja V, Nuyujukian P, Foster JD, Fan JM, Kaufman MT, Churchland MM, Rivera-Alvidrez Z, Cunningham JP, Ryu SI, Shenoy KV (2011) Long-term stability of neural prosthetic control signals from silicon cortical arrays in rhesus macaque motor cortex. J Neural Eng 8:045005. CrossRef Medline

Cohen MR, Maunsell JH (2009) Attention improves performance primarily by reducing interneuronal correlations. Nat Neurosci 12:1594–1600. CrossRef Medline

Cohen MR, Maunsell JH (2010) A neuronal population measure of attention predicts behavioral performance on individual trials. J Neurosci 30:15241–15253. CrossRef Medline

Cohen MR, Maunsell JH (2011) Using neuronal populations to study the mechanisms underlying spatial and feature attention. Neuron 70:1192–1204. CrossRef Medline

Connor CE, Hsiao SS, Phillips JR, Johnson KO (1990) Tactile roughness: neural codes that account for psychophysical magnitude estimates. J Neurosci 10:3823–3836. Medline

Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. Curr Opin Neurobiol 17:140–147. CrossRef Medline

Desimone R, Albright TD, Gross CG, Bruce C (1984) Stimulus-selective properties of inferior temporal neurons in the macaque. J Neurosci 4:2051–2062. Medline

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341. CrossRef Medline

DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? Neuron 73:415–434. CrossRef Medline

Engel AK, Fries P, Singer W (2001) Dynamic predictions: oscillations and synchrony in top-down processing. Nat Rev Neurosci 2:704–716. CrossRef Medline

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1:1–47. Medline

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312–316. CrossRef Medline

Freeman J, Ziemba CM, Heeger DJ, Simoncelli EP, Movshon JA (2013) A functional and perceptual signature of the second visual area in primates. Nat Neurosci 16:974–981. CrossRef Medline

Frey BJ, Dueck D (2007) Clustering by passing messages between data points. Science 315:972–976. CrossRef Medline

Gallant JL, Connor CE, Rakshit S, Lewis JW, Van Essen DC (1996) Neural responses to polar, hyperbolic, and Cartesian gratings in area V4 of the macaque monkey. J Neurophysiol 76:2718–2739. Medline

Hegdé J, Van Essen DC (2000) Selectivity for complex shapes in primate visual area V2. J Neurosci 20:RC61. Medline

Holmes EJ, Gross CG (1984) Effects of inferior temporal lesions on discrimination of stimuli differing in orientation. J Neurosci 4:3063–3068. Medline

Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast readout of object identity from macaque inferior temporal cortex. Science 310:863–866. CrossRef Medline

Hyvarinen J, Sakata H, Talbot WH, Mountcastle VB (1968) Neuronal coding by cortical cells of the frequency of oscillating peripheral stimuli. Science 162:1130–1132. CrossRef Medline

Issa EB, DiCarlo JJ (2012) Precedence of the eye region in neural processing of faces. J Neurosci 32:16666–16682. CrossRef Medline

Johnson KO, Hsiao SS, Yoshioka T (2002) Neural coding and the basic law of psychophysics. Neuroscientist 8:111–121. CrossRef Medline

Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352–355. CrossRef Medline

Kersten D, Mamassian P, Yuille A (2004) Object perception as Bayesian inference. Annu Rev Psychol 55:271–304. CrossRef Medline

Kobatake E, Tanaka K (1994) Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. J Neurophysiol 71:856–867. Medline

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–1141. CrossRef Medline

Lazebnik S, Schmid C, Ponce J (2009) Spatial pyramid matching. Object Categorization: Computer and Human Vision Perspectives 3:4.

Li N, DiCarlo JJ (2008) Unsupervised natural experience rapidly alters invariant object representation in visual cortex. Science 321:1502–1507. CrossRef Medline

Liu S, Dickman JD, Newlands SD, DeAngelis GC, Angelaki DE (2013) Reduced choice-related activity and correlated noise accompany perceptual deficits following unilateral vestibular lesion. Proc Natl Acad Sci U S A 110:17999–18004. CrossRef Medline

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vision 60:91–110. CrossRef

Luna R, Hernández A, Brody CD, Romo R (2005) Neural codes for perceptual discrimination in primary somatosensory cortex. Nat Neurosci 8:1210–1219. CrossRef Medline

Merigan WH (1976) The contrast sensitivity of the squirrel monkey (*Saimiri sciureus*). Vision Res 16:375–379. CrossRef Medline

Meyers EM, Freedman DJ, Kreiman G, Miller EK, Poggio T (2008) Dynamic population coding of category information in inferior temporal and prefrontal cortex. J Neurophysiol 100:1407–1419. CrossRef Medline

Mitchell JF, Sundberg KA, Reynolds JH (2009) Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron 63:879–888. CrossRef Medline

Mountcastle VB, Talbot WH, Sakata H, Hyvärinen J (1969) Cortical neuronal mechanisms in flutter-vibration studied in unanesthetized monkeys: neuronal periodicity and frequency discrimination. J Neurophysiol 32:452–484. Medline

Mutch J, Lowe DG (2008) Object class recognition and localization using sparse features with limited receptive fields. Int J Comput Vision 80:45–57. CrossRef

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL (2009) Bayesian reconstruction of natural images from human brain activity. Neuron 63:902–915. CrossRef Medline

Newsome WT, Britten KH, Movshon JA (1989) Neuronal correlates of a perceptual decision. Nature 341:52–54. CrossRef Medline

Oliva A, Torralba A (2006) Building the gist of a scene: the role of global image features in recognition. Prog Brain Res 155:23–36. CrossRef Medline

Op de Beeck H, Wagemans J, Vogels R (2001) Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat Neurosci 4:1244–1252. CrossRef Medline

Pagan M, Urban LS, Wohl MP, Rust NC (2013) Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. Nat Neurosci 16:1132–1139. CrossRef Medline

Parker AJ, Newsome WT (1998) Sense and the single neuron: probing the physiology of perception. Annu Rev Neurosci 21:227–277. CrossRef Medline

Pasupathy A, Connor CE (1999) Responses to contour features in macaque area V4. J Neurophysiol 82:2490–2502. Medline

Pasupathy A, Connor CE (2002) Population coding of shape in area V4. Nat Neurosci 5:1332–1338. CrossRef Medline

Pinto N, Cox DD, DiCarlo JJ (2008) Why is real-world visual object recognition hard? PLoS Comput Biol 4:e27. CrossRef Medline

Pinto N, Barhomi Y, Cox DD, DiCarlo JJ (2011) Comparing state-of-the-art visual features on invariant object recognition tasks. In: IEEE Workshop on Applications of Computer Vision (Kona, HI).

Quiroga RQ, Nadasdy Z, Ben-Shaul Y (2004) Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput 16:1661–1687. CrossRef Medline

Rajalingham R, Schmidt K, DiCarlo JJ (2015) Comparison of object recognition behavior in human and monkey. J Neurosci 35:12127–12136. CrossRef

Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025. CrossRef Medline

Rust NC, DiCarlo JJ (2010) Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. J Neurosci 30:12978–12995. CrossRef Medline

Rust NC, DiCarlo JJ (2012) Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. J Neurosci 32:10170–10182. CrossRef Medline

Serre T, Oliva A, Poggio T (2007) A feedforward architecture accounts for rapid categorization. Proc Natl Acad Sci U S A 104:6424–6429. CrossRef Medline

Shadlen MN, Britten KH, Newsome WT, Movshon JA (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. J Neurosci 16:1486–1510. Medline

Sheinberg DL, Logothetis NK (1997) The role of temporal cortical areas in perceptual organization. Proc Natl Acad Sci U S A 94:3408–3413. CrossRef Medline

Sigala N, Gabbiani F, Logothetis NK (2002) Visual categorization and object representation in monkeys and humans. J Cogn Neurosci 14:187–198. CrossRef Medline

Sugase Y, Yamane S, Ueno S, Kawano K (1999) Global and fine information coded by single neurons in the temporal visual cortex. Nature 400:869–873. CrossRef Medline

Tanaka K (1993) Neuronal mechanisms of object recognition. Science 262:685–688. CrossRef Medline

Tanaka K (1997) Mechanisms of visual object recognition: monkey and human studies. Curr Opin Neurobiol 7:523–529. CrossRef Medline

Tarr MJ, Bülthoff HH (1998) Image-based object recognition in man, monkey and machine. Cognition 67:1–20. CrossRef Medline

Tjan BS, Legge GE (1998) The viewpoint complexity of an object-recognition task. Vision Res 38:2335–2350. CrossRef Medline

Tsao DY, Freiwald WA, Tootell RB, Livingstone MS (2006) A cortical region consisting entirely of face-selective cells. Science 311:670–674. CrossRef Medline

Verhoef BE, Vogels R, Janssen P (2012) Inferotemporal cortex subserves three-dimensional structure categorization. Neuron 73:171–182. CrossRef Medline

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc Natl Acad Sci U S A 111:8619–8624. CrossRef Medline

Yoshioka T, Gibb B, Dorsch AK, Hsiao SS, Johnson KO (2001) Neural coding mechanisms underlying perceived roughness of finely textured surfaces. J Neurosci 21:6905–6916. Medline

Zoccolan D, Oertelt N, DiCarlo JJ, Cox DD (2009) A rodent model for the study of invariant visual object recognition. Proc Natl Acad Sci U S A 106:8748–8753. CrossRef Medline

Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370:140–143. CrossRef Medline