Neuron

Article

Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits

Highlights

- Focal inactivation of IT resulted in reliable deficits in core object recognition
- Inactivating different IT sites resulted in different patterns of object deficits
- Deficit patterns were topographically organized over the cortical surface
- Deficit patterns were predicted by each IT site's neuronal object discriminability

Authors

Rishi Rajalingham, James J. DiCarlo

Correspondence

rishir@mit.edu (R.R.),
dicarlo@mit.edu (J.J.D.)

In Brief

Rajalingham and DiCarlo show that inactivating millimeter-scale IT subregions results in selective object recognition deficits, providing direct evidence for a causal role of IT in this behavior. Inactivating different subregions resulted in different deficit patterns, revealing an underlying topographical organization.

Rajalingham & DiCarlo, 2019, Neuron *102*, 1–13
April 17, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.neuron.2019.02.001

CellPress

Please cite this article in press as: Rajalingham and DiCarlo, Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits, Neuron (2019), https://doi.org/10.1016/j.neuron.2019.02.001

Neuron
Article

Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits

Rishi Rajalingham[1,*] and James J. DiCarlo[1,2,*]
[1]McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Lead Contact
*Correspondence: rishir@mit.edu (R.R.), dicarlo@mit.edu (J.J.D.)
https://doi.org/10.1016/j.neuron.2019.02.001

SUMMARY

Extensive research suggests that the inferior temporal (IT) population supports visual object recognition behavior. However, causal evidence for this hypothesis has been equivocal, particularly beyond the specific case of face-selective subregions of IT. Here, we directly tested this hypothesis by pharmacologically inactivating individual, millimeter-scale subregions of IT while monkeys performed several core object recognition subtasks, interleaved trial-by trial. First, we observed that IT inactivation resulted in reliable contralateral-biased subtask-selective behavioral deficits. Moreover, inactivating different IT subregions resulted in different patterns of subtask deficits, predicted by each subregion's neuronal object discriminability. Finally, the similarity between different inactivation effects was tightly related to the anatomical distance between corresponding inactivation sites. Taken together, these results provide direct evidence that the IT cortex causally supports general core object recognition and that the underlying IT coding dimensions are topographically organized.

INTRODUCTION

Primate core visual object recognition—the ability to rapidly recognize objects in the central 10 degrees in spite of naturally occurring identity-preserving image variability—is thought to rely on the ventral visual stream, a hierarchy of visual cortical areas (DiCarlo et al., 2012). Decades of research suggest that the inferior temporal (IT) cortex, the highest level of the ventral stream hierarchy, is a necessary part of the brain's neural network that underlies core recognition behavior (Logothetis and Sheinberg, 1996; Tanaka, 1996; Rolls, 2000; DiCarlo et al., 2012). For example, it has been shown that parallel linear object discriminants acting on the IT population not only match overall primate behavioral performance (Hung et al., 2005; Zhang et al., 2011) but also predict primate behavioral patterns (Sheinberg and Logothetis, 1997; Op de Beeck et al., 2001; Majaj et al., 2015), showing that IT is a tight neural correlate of primate recognition behavior. Quantitative versions of such experiments have proposed downstream neurally mechanistic models that successfully link IT population activity to behavior (Majaj et al., 2015), mechanisms that appear to accurately generalize to all core object recognition subtasks (e.g., "car" versus "not car," "face" versus "not face," etc.). While these experiments are consistent with the hypothesis that IT is a necessary node in the neural network supporting core object recognition behavior, they might also be epiphenomenal (Katz et al., 2016; Liu and Pack, 2017). To directly infer the causal role of IT in this behavior, it is necessary to bring the IT activity under more direct experimenter control (e.g., via the application of pharmacological agents into IT to silence neurons, etc.) while measuring behavior.

To date, the most successful direct IT manipulations in the context of object recognition have targeted millimeter-scale clusters of face-selective neurons in IT (Afraz et al., 2006, 2015; Moeller et al., 2017; Sadagopan et al., 2017). These studies suggest that neurons in these IT subregions are necessary for at least some basic- and subordinate-level face recognition behaviors. Beyond this domain, a notable study by Verhoef et al. (2012) found that manipulation of clusters of 3D-structure-preferring neurons in IT influenced the categorization of 3D stimuli as convex or concave. However, results from direct manipulations of IT in general visual object recognition behavior have been equivocal at best. Lesions of IT sometimes suggest the necessity of IT and visual behaviors (Cowey and Gross, 1970; Manning, 1972; Holmes and Gross, 1984; Weiskrantz and Saunders, 1984; Biederman et al., 1997; Buffalo et al., 2000), but the resulting behavioral deficits are often contradictory (often with no lasting visual deficits; Dean, 1974; Huxlin et al., 2010) and surprisingly modest even for large-scale bilateral removal of IT (e.g., 10%–15% drop in performance when complete loss of performance would have been 40%) (Horel et al., 1987; Matsumoto et al., 2016). Thus, it is still unclear whether IT is a necessary node in supporting general core object recognition behavior. Moreover, even if the IT cortex is indeed necessary for all core object recognition subtasks, it is unclear whether that assumed causal role is spatially organized. For example, the current literature on monkey IT is consistent with the hypothesis that every square millimeter of the IT cortex outside of the fMRI-defined face patches is equally involved in all (non-face) object

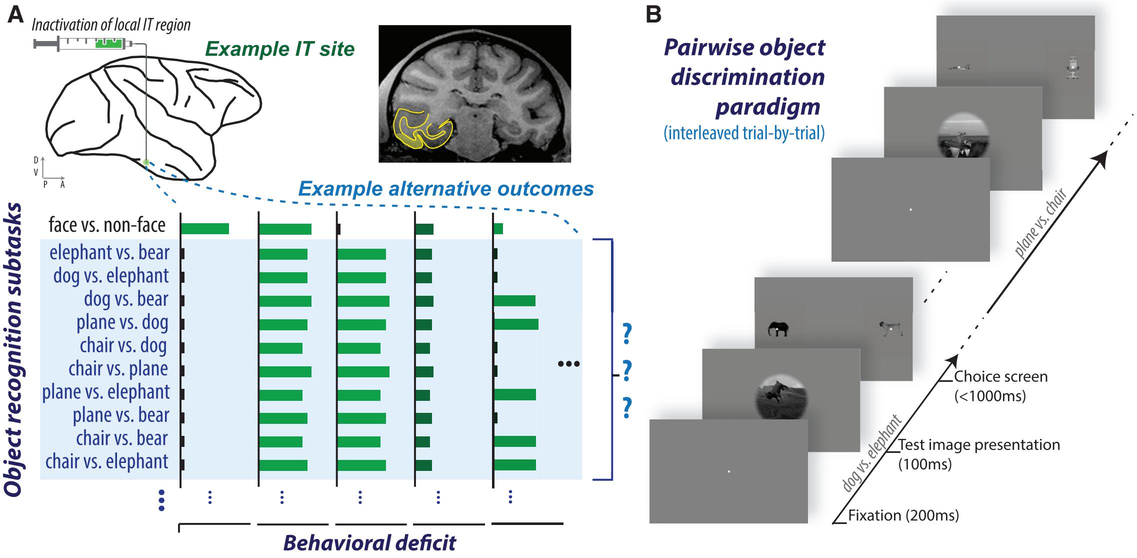

Figure 1. Experimental Design and Hypothesis

(A) Schematic of experiment. It is still unclear whether IT is necessary for general core object recognition behavior, and, moreover, whether any such causal role is functionally specific at the millimeter scale. To investigate this, we reversibly inactivated individual arbitrarily sampled millimeter-scale regions of IT via local injection of muscimol while monkeys performed a battery of pairwise core object recognition subtasks (listed, highlighted in blue), interleaved trial by trial (see B). The inset shows an MRI coronal slice highlighting the ventral surface of IT, the region targeted for inactivation experiments. These subtasks were pseudo-randomly selected from the large set of pairwise discriminations that animals were previously trained on, with the explicit goal of testing "arbitrary" basic-level object recognition subtasks. Bar plots outline alternative possible outcomes corresponding to different patterns of behavioral deficits from such inactivations, varying in subtask selectivity from highly specialized (far left panel, exhibiting deficits only for face versus non-face discriminations), to largely uniform (middle three panels, exhibiting equal deficits on all discrimination subtasks, or all non-face discrimination subtasks), to relatively subtask-selective (far right panel, exhibiting deficits on some but not all discrimination subtasks).

(B) Behavioral paradigm. Each trial was initiated when the monkey acquired and held its gaze on a central fixation point for 200 ms, after which a test image (8×8 degrees of visual angle) appeared at the center of gaze for 100 ms. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right. One of these two objects was always the same as the object that generated the test image (i.e., the correct choice), and its location (left or right) was randomly chosen on each trial. The monkey was allowed to freely view the choice images for up to 1,000 ms and indicated its final choice by holding fixation over the selected image for 700 ms. A juice reward was delivered immediately after each correct trial. Note that we refer to each pairwise object discrimination (averaged over all test images) as a "discrimination subtask" and the trials for all such subtasks (6–10 subtasks, see STAR Methods) were pseudo-randomly interleaved trial by trial.

discriminations, with some authors implicitly arguing for that hypothesis (Tsao and Livingstone, 2008; Kanwisher, 2010).

To investigate these open questions, we reversibly inactivated neurons in individual, arbitrarily sampled millimeter-scale regions of the ventral surface of IT via local injection of muscimol (a single injection of 1 μl of muscimol, corresponding to strong neural suppression in a volume ~ 2.5 mm in diameter, centered at the injection site; Arikan et al., 2002) while monkeys performed a battery of pairwise core object discrimination subtasks, interleaved trial by trial. This paradigm allowed us not only to test the aforementioned IT-to-behavior linking hypotheses directly (Majaj et al., 2015) but also to characterize the causal role of each inactivation IT site via a *pattern* of deficits over object recognition subtasks.

Our results show that inactivation of even single, millimeter-scale regions of IT resulted in reliable contralateral-biased behavioral deficits. Interestingly, these deficits were highly selective over core object recognition subtasks—inactivating a small region of IT produced deficits in only a subset of such subtasks, and inactivating different such regions resulted in different patterns of object recognition deficits. Furthermore, the effect of inactivation was topographically organized in that the pattern of behavioral deficit (i.e., the pattern over subtasks) was most similar at anatomically neighboring injection sites. We also found that each pattern of subtask deficit was well predicted by the object discriminability of the local region's neuronal activity. Taken together these results demonstrate the necessity of the IT cortex for a wide range of general core object recognition behaviors and reveal that—even outside of face patches—the IT cortex has behaviorally critical topographic organization of visual features. These findings are consistent with and suggested by prior physiology work (Wang et al., 1998; Tsunoda et al., 2001; Kreiman et al., 2006 for sub-millimeter columnar organization; Lafer-Sousa and Conway, 2013; Conway, 2018 for broad spatial organization of IT), but, to our knowledge, this is the first demonstration of a topographically organized causal role of IT in general core object recognition.

RESULTS

Our primary goal was to ask whether IT causally supports object recognition, and whether any such causal role is functionally specific at the millimeter scale, as schematized in Figure 1A.

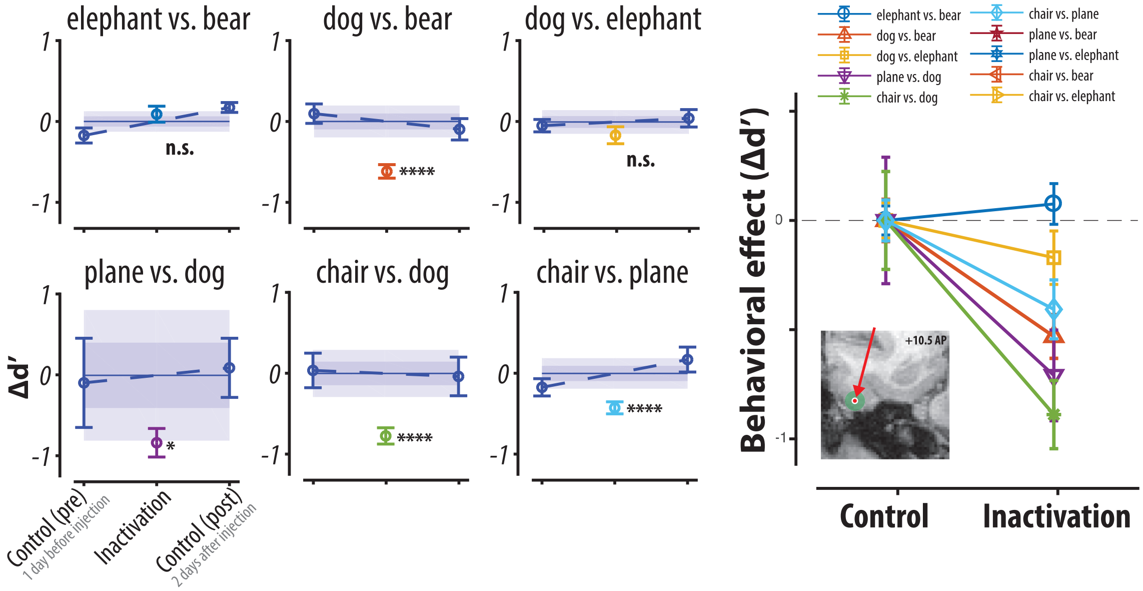

Figure 2. Example Inactivation Experiment

Example inactivation experiment. Behavioral performance (mean) for each of six subtasks over the three condition (pre-control, inactivation, and post-control; see STAR Methods) is shown at left. Data are shown as behavioral performance relative the average of pre- and post-control performances (see STAR Methods) (bars show SEM obtained by bootstrap resampling over trials). The location of the injection site ("inactivation" condition) for this experiment is shown in the right panel. The dark and light shaded areas correspond to one and two SEM respectively of this control. For this site, we observed a strong and significant deficit for some subtasks (chair versus dog, chair versus plane, and dog versus bear) but not others (elephant versus bear or dog versus elephant). The data on the left are summarized relative to the average control performance (mean ± SEM over trials) in the right panel. The inset shows an MRI coronal slice highlighting the anatomical location and extent of the inactivation.

To do this, we reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of the ventral surface of IT via injection of 1 μl of muscimol while monkeys performed a battery of pairwise core object discrimination subtasks. For this volume of injection, we expect strong neural suppression in a volume ∼2.5 mm in diameter centered at the injection site (Arikan et al., 2002). Figure 1B shows the behavioral paradigm used for testing monkeys' core object recognition behavior. In this work, the battery consisted of 6 (Monkeys 1 and 2) or 10 (Monkey 2 only) pairwise core object discrimination subtasks between five objects, interleaved trial by trial (see Figure 1A for subtask list). These subtasks were pseudorandomly selected from the large set of pairwise discriminations that animals were previously trained on, with the explicit goal of testing "arbitrary" basic-level object recognition subtasks. To enforce true object recognition (rather than image matching), stimuli consisted of naturalistic synthetic images of 3D objects rendered under high view uncertainty (see Figure S1A for example images), and the monkey subjects were required to generalize to new images in each subtask (as we have previously shown they readily do; Rajalingham et al., 2015).

Figure 2 shows the behavioral data for an example inactivation experiment in Monkey 1 for each of six pairwise discrimination subtasks. Each panel on the left shows the relative behavioral performance (mean ± SEM, obtained by bootstrap resampling over trials) for a given pairwise subtask for each of three consecutive behavioral sessions (pre-inactivation control, inactivation, and post-inactivation control; see STAR Methods). Performance on each subtask is shown relative to the average performance on that subtask over the pre- and post-control sessions; this definition of control behavior aims to be robust to natural variability in performance across behavioral sessions (see STAR Methods). The dark- and light-shaded areas correspond to one and two SEM of this measure (computed over trials), respectively. We observed a strong and significant deficit due to inactivation for some subtasks (i.e., chair versus dog, chair versus plane, and dog versus bear) but not others (elephant versus bear or dog versus elephant). The resulting pattern of behavioral deficits (i.e., the deficit pattern over subtasks) for this one example inactivation site in IT is shown on the right panel, with the corresponding anatomical location shown in the inset.

Summary of Behavioral Deficits

Figure 3A shows the behavioral deficits for all inactivation sites and all subtasks in both monkeys as a scatter of control performance versus inactivation performance (n = 25 sites, n = 182 subtasks × sites). Considering all the subtasks together, we observed a significant decrease in performance (i.e., inactivation lower than control), corresponding to the predominance of points under the unity line in Figure 3; on average, this amounted to a global deficit of $\langle \Delta d' \rangle = -0.2 \pm 0.02$ ($p < 10^{-15}$, one-tailed exact test; see Figure 3B, red bar under "global deficit"). Additionally, we observed global changes in balanced accuracy $(\mu = -2\%, p < 10^{-5})$ and choice bias $(\mu = -0.23, p < 10^{-2})$

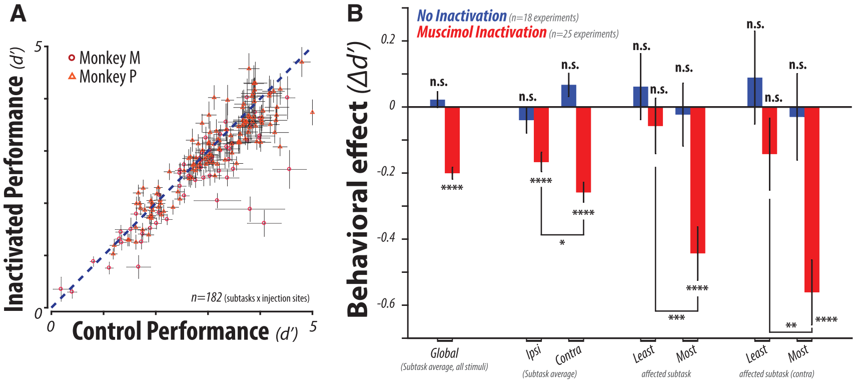

Figure 3. Summary of Inactivation Effects

(A) Behavioral deficits for all IT inactivation sites and all subtasks in both monkeys as a scatter of control performance and inactivation performance. Note the on-average decrease in performance corresponding to predominance of points under the unity line (dashed line).

(B) Summary of behavioral deficits when grouping the subtasks and subtask images in different ways. Red bars show the magnitude of inactivation deficit (relative to control) for each grouping. From left to right, these groupings are: all images and all subtasks ("Global"), ipsilateral/contralateral object images for all subtasks ("Ipsi" and "Contra"), the least/most affected subtask at each site ("Least" and "Most") selected on held out data, and contralateral object images for the least/most affected subtask at each site ("Most Contra") selected on held out data. Blue bars correspond to otherwise identical experiments but without muscimol inactivation (control experiments).

(Figure S2A). Here, we focus our analyses with respect to changes in sensitivity (d') for principled reasons (see STAR Methods).

Consistent with the known lateralization of IT (Op de Beeck and Vogels, 2000), this deficit was more pronounced for images in which the center of the target object was contralateral to the injection hemisphere ($\langle \Delta d' \rangle = -0.26 \pm 0.03$, $p < 10^{-15}$) than for images with ipsilateral object centers ($\langle \Delta d' \rangle = -0.17 \pm 0.03$, $p < 10^{-11}$), and this difference was significant ($p = 0.0128$, one-tailed exact test; ipsi versus contra). Note that all images were presented foveally, spanning -4° to 4° of both azimuth and elevation, and average object size was $\sim 3.5^\circ$.

Next, we asked whether the inactivation deficits were subtask-specific. To examine this, we compared the magnitude of behavioral deficits between the least-affected and most-affected subtasks for each inactivation site. Crucially, to avoid any selection bias, these subtasks were selected from held-out data: we split our data into two disjoint halves of trials, selected the least- and most-affected subtasks per inactivation site from one split half, and examined the corresponding deficits on these selected subtasks in the second split half (thus, the expected value of the difference in deficits between the most and least affected subtask is zero under the null hypothesis; see STAR Methods). Using this procedure, we observed a large significant behavioral deficit for the most affected subtask ($\langle \Delta d' \rangle = -0.44 \pm 0.08$, $p < 10^{-15}$) but not for the least-affected subtask ($\langle \Delta d' \rangle = -0.06 \pm 0.08$, $p = 0.27$), and the difference was significant ($p < 10^{-3}$; see Figure 3B). Finally, we observed even larger subtask-selective deficits when restricting to contralateral objects (as described above), with a similar significant difference between the most- and least-affected subtasks ($\langle \Delta d' \rangle = -0.56 \pm 0.10, -0.14 \pm 0.11$ for most- and least-affected subtasks, respectively; $p < 10^{-2}$).

For each of the analyzed conditions, we observed no significant behavioral deficits on otherwise identical experiments without muscimol inactivation ($p > 0.05$; Figure 3B, blue bars). Furthermore, the patterns of deficits across these analyzed conditions were similar for both animals (Figure S3A). In summary, inactivation of local regions of IT resulted in highly reliable behavioral deficits, which were selective over visual space (i.e., contralateral-biased) and selective over different core object recognition subtasks. For the remaining analyses, we focus on "contralateral stimuli" (i.e., images in which the center of the target object was contralateral to the injection hemisphere) to characterize the behavioral effects of focal IT inactivation.

Figure 4 shows the deficit patterns for each of the 25 individual inactivation sites, formatted as in Figure 2. We qualitatively observe that behavioral performance on one or more, but not all, object discrimination subtasks is typically reduced by inactivation of each IT site, and that the specific subtask(s) affected are different at different IT sites. The average behavioral effect over subtasks was negative (consistent with a behavioral *deficit*) for a significant proportion of individual sites (see Figure 4, Venn diagram; significantly negative effect: 40%,$p = 0.003$; χ^2 test for proportions). For a small and non-significant proportion of sites, we observed a positive average behavioral effect over subtasks (significantly positive effect: 4%,$p = 0.86$, χ^2 test); we speculate that this could reflect random experimental variability. Together, these results suggest that inactivating different millimeter-scale regions of primate IT results in deficits in different core object recognition subtasks (i.e., different patterns of deficits). This inference is directly and quantitatively tested in the following analyses.

Task-Selectivity of Deficits

Figure 5A shows the subtask deficit patterns for each of the 25 inactivation sites as a heatmap. Each column corresponds to the deficit pattern over subtasks from inactivating an individual IT site, normalized to a fixed color scale (0,1); brighter colors correspond to larger relative subtask deficits. Consistent with the inferred subtask-selectivity from Figure 3B, we observed that each inactivation resulted in a non-uniform behavioral deficit pattern. This non-uniformity was quantified via a sparsity index (*SI*; see STAR Methods), which has a value of 0 for perfectly uniform deficit patterns (i.e., where each IT subregion is equally necessary for all subtasks), and a value of 1 for a perfectly subtask-specialized (or "one-hot") deficit pattern (i.e., where each subregion is necessary for just one of the tested subtasks). We

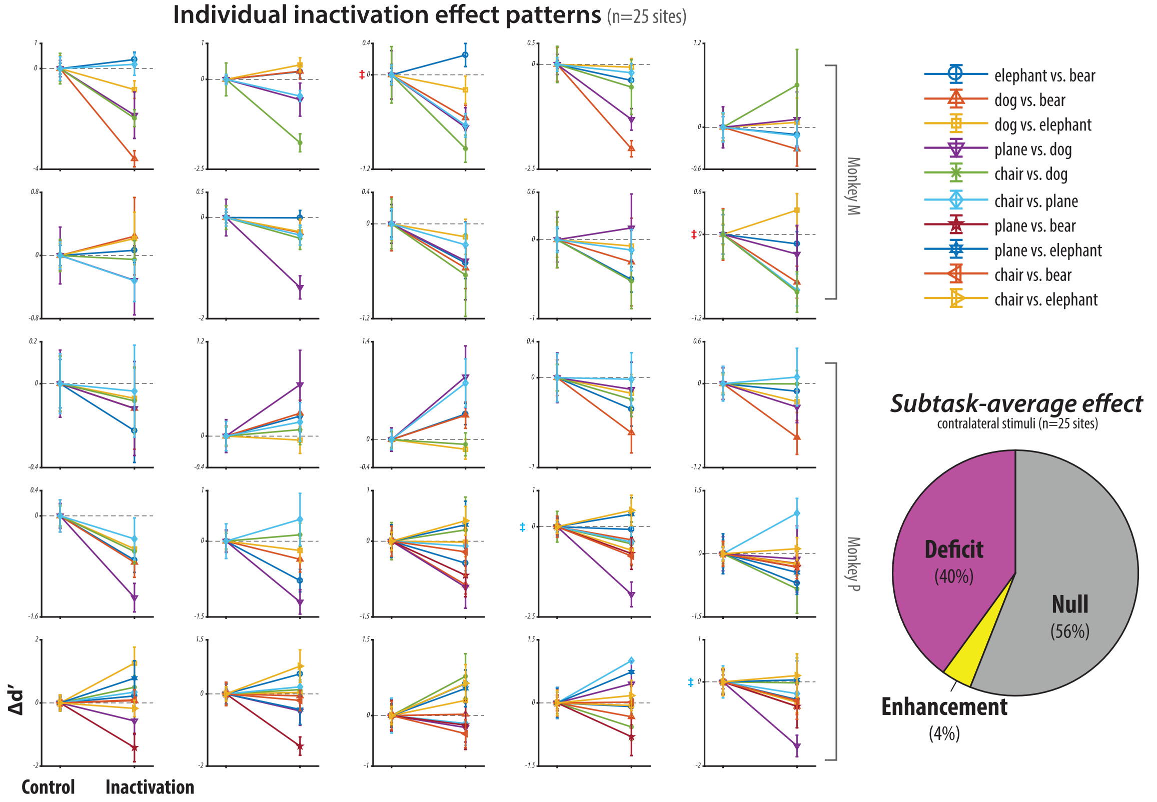

Figure 4. Individual Inactivation Deficit Patterns

Individual inactivation deficit patterns. Each of the 25 panels shows the inactivation pattern, formatted as in Figure 2, for each of the 25 individual inactivation sites. The most-closely neighboring site pairs in each monkey are indicated with colored daggers (‡). Sites were categorized based on the sign (positive/negative) of the average behavioral effect over subtasks, with significance of each site assessed by a two-tailed exact test (at p<0.05); the Venn on the bottom right diagram shows the proportion of sites in each category.

observed that inactivation of local regions in IT led to highly non-uniform deficit patterns, on average ($SI(\delta) = 0.71 \pm 0.05$; mean ± SEM over sites, see Figure 3D).

To ground this empirical *SI* value, we estimated the corresponding *SI* distributions for different simulated behavioral deficit patterns with varying degrees of non-uniformity across subtasks. These simulated deficit patterns were obtained via random permutations of our data, varying only the proportion of affected subtasks (see STAR Methods). Crucially, the simulated sparseness distributions preserved a number of key sources of variance—including the number of sites, the number of subtasks for each site, the performance on each subtask for each site, and the average performance deficit (across subtasks) for each site—because all these sources of variance were fixed for the empirical and simulated sparseness estimates, and the random shuffling was done after computing the behavioral deficits. Figure 5B shows the *SI* distributions expected from behavioral deficits of varying degrees of non-uniformity (i.e., with $10\%, 25\%, \ldots, 100\%$ of subtasks affected). We observe that the empirically observed subtask selectivity is significantly greater than expected from a uniform deficit ($p < 10^{-15}$; relative to simulated 100% affected, i.e., uniform) but significantly less than expected from a highly sparse deficit pattern ($p < 10^{-2}$; relative to simulated 10% affected). Indeed, the observed *SI* estimates correspond to simulation of deficits on $\sim 25\%$ of tested subtasks.

Importantly, this non-uniformity does not simply reflect non-uniformity in the behavioral difficulty across subtasks. Indeed, normalizing each deficit pattern by the behavioral difficulty pattern resulted in normalized deficit patterns that were not significantly correlated with subtask difficulty ($r = 0.06$, $p = 0.39$) and significantly non-uniform as quantified by sparsity ($SI(\delta_n) = 0.74 \pm 0.06$; $p < 10^{-5}$, relative to simulated uniform). This is also clear from Figure 5A, which shows that inactivation of different sites led to different deficit weight patterns (left panel). Accordingly, the deficits were relatively evenly distributed over the subtasks, as reflected by the approximate uniformity (except for one subtask, plane versus bear) of the average deficit pattern over all sites (Figure 5A, rightmost bar). Together, these results indicate that the non-uniformity of subtask deficits is not tied to specific subtasks.

Additionally, we tested whether the deficits were evenly distributed over the five objects across all inactivation sites. To do this, we computed the pattern of deficits over objects (one-versus-all performance; see STAR Methods) for each inactivation site, then we estimated the average normalized deficit and the probability that each of the five tested objects corresponds to the most affected object per inactivation site. As

Please cite this article in press as: Rajalingham and DiCarlo, Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits, Neuron (2019), https://doi.org/10.1016/j.neuron.2019.02.001

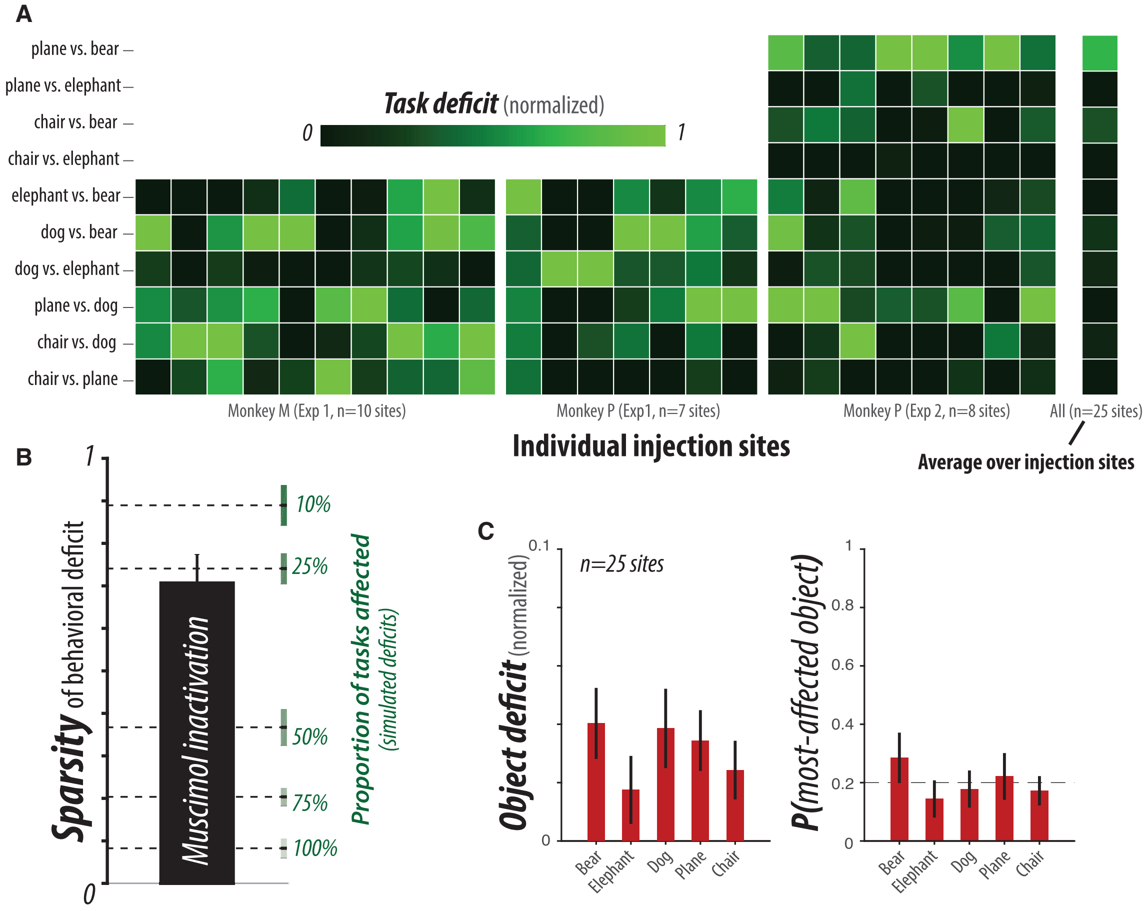

Figure 5. Task-Selectivity of Inactivation Deficit Patterns

(A) The heatmap shows the subtask deficits for each of the 25 inactivation sites, with brighter colors corresponding to larger relative subtask deficits, highlighting that inactivation of each IT site resulted in a different, relatively sparse, pattern of behavioral deficit. The average deficit pattern over all inactivation sites (right column) is largely uniform, suggesting that IT, as a whole, is approximately equally involved in each discrimination subtask.

(B) The black bar shows the sparsity (see STAR Methods) of the behavioral subtask deficits over all sites (mean ± SEM over sites). To provide calibration, green lines show the sparsity values that occur under simulations in which we varied the proportion of truly affected subtasks and used identical sampling noise as our data (see STAR Methods). Together, these results suggest that inactivation of a single 2.5-mm-diameter region of IT affects 25% of core object discrimination subtasks, on average.

(C) Distribution of deficits over objects. The left panel shows the normalized deficit $\left(-\left(\Delta d'/d'\right)\right)$, with respect to a one-versus-all behavioral metric, over each of the five tested objects (mean ± SEM, over trials). The right panel shows the probability of each of the five objects to be the most-affected object per inactivation site (mean ± SEM, over trials).

shown in Figure 5C, the average normalized deficit was not significantly different across objects, and the estimated probabilities were not significantly different from chance (20%), consistent with relatively evenly distributed deficits over objects, across inactivation sites.

Tissue Selectivity of Deficits

Inactivating different anatomical regions of IT resulted in different patterns of subtask deficits. To directly test this tissue selectivity, we compared the inactivation deficit patterns between pairs of IT sites. Pairwise deficit pattern similarity was quantified using a noise-adjusted correlation ($\tilde{\rho}$; see STAR Methods). We considered all pairs of inactivation sites, measured within the same animal and image-set, where the inactivation deficit patterns of both sites had split-half internal reliability greater than a threshold θ ($n = 62$ pairs for $\theta =$ 0.1, but results did not significantly depend on the choice of the threshold θ). We measured the dependence of pairwise deficit similarity on the anatomical distance between the inactivation sites, where anatomical distance (d) was computed as the Euclidean distance between the injection site locations estimated via high-resolution microfocal stereo X-ray reconstruction (see STAR Methods). First, we observed that inactivation deficits are highly replicable across experiments: the noise-adjusted correlation between behavioral deficit patterns of neighboring inactivation sites was near ceiling ($\tilde{\rho} = 0.92 \pm 0.03$ for $d < 1$mm, mean ± SEM over site pairs; Figure 6A). In other words, we infer that repeated inactivation of the "same" anatomical site (within a small margin of error) leads to reproducible behavior deficit patterns. This is evidenced by the similar deficit patterns for the most-closely neighboring pair of inactivation sites in each monkey (highlighted with

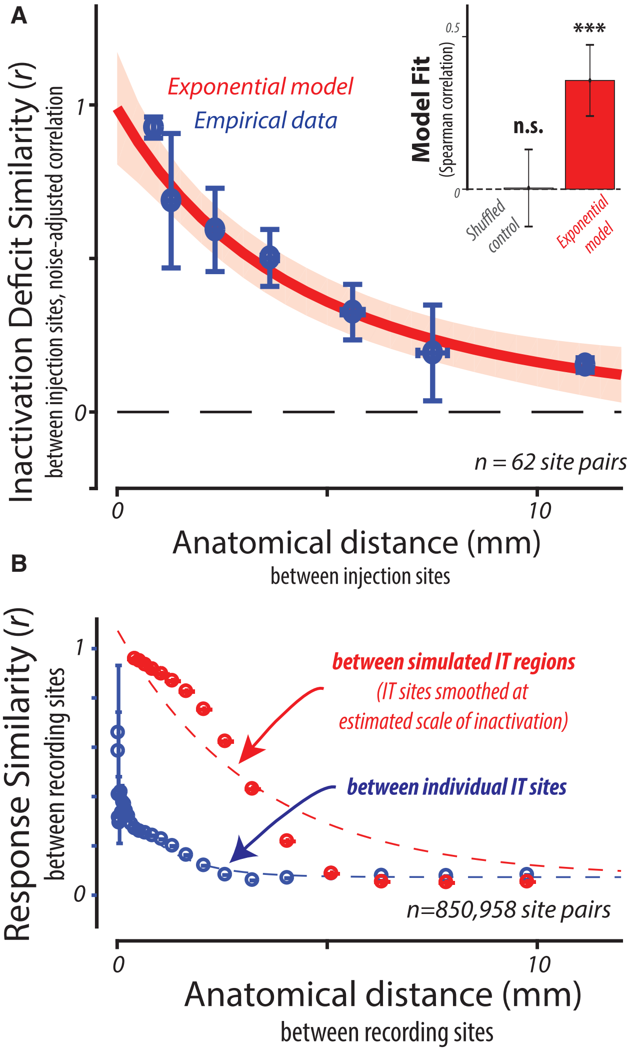

Figure 6. Tissue-Selectivity of Inactivation Effects

(A) Topographical organization. Similarity in behavioral deficit patterns between pairs of IT injection sites (quantified as noise-adjusted correlation, y axis) as a function of the anatomical distance between each pair of sites (x axis). Empirical data are shown as the mean (± SEM) of all pairs of sites in logarithmically-spaced bins of tissue distance (blue points). Note that the pattern of inactivation-induced behavioral effects is highly replicable in that we observe very high correlation of effects for repeated experiments at or very near the originally tested site (near 0 on the x axis). The similarity between any two inactivation deficits was monotonically related to their anatomical distance, and a simple exponential model significantly explained this relationship (see inset). Note that the model correlation was estimated from the raw empirical data (over all 62 site pairs) and did not depend on the logarithmically spaced binning.

(B) Response similarity between neighboring neurons in IT cortex, computed from a previously recorded large-scale high-resolution neurophysiological dataset (Issa et al., 2013b), highlighting the sub-millimeter scale organization of IT responses (blue). Expected similarity of mm-scale regions of IT, obtained by smoothing neural responses at the scale of muscimol inactivation (red).

daggers (‡) in Figure 4). Importantly, this similarity between inactivation deficit patterns did not reflect changes in behavior over time (e.g., from subtask exposure), as shown in Figure S4B. Further, we observed that this similarity between the inactivation deficits of two injection sites was monotonically related to the anatomical distance between them (Figure 6A, see Figure S3B for qualitatively similar plots in each animal separately). A simple exponential decay model (half-max-full-width $HMFW = 3.29 \pm 1.19$mm) significantly explained this relationship ($R^2 = 0.36 \pm 0.12$, $p < 10^{-3}$). Note that the model correlation was estimated from the raw empirical data (i.e., all 62 site pairs), and did not depend on the logarithmically spaced binning. We verified that this model correlation is not expected by chance by fitting the model on randomly shuffled data ($R^2 = 0.00 \pm 0.13$, $p = 0.50$).

In an attempt to quantitatively relate this observed topographical organization of deficits to the underlying spatial organization of IT, we computed the similarity in image-driven response patterns of pairs of IT neurons, obtained from a previously recorded large-scale high-resolution neurophysiological dataset (Issa et al., 2013b). We observed strong spatial organization at a sub-millimeter scale in IT (see Figure 6B, blue lines). From these data, we simulated the expected similarity of regions of IT at the mm-scale expected from our muscimol inactivation (Arikan et al., 2002) (see Figure 6B, red lines) by spatially convolving neural responses with a boxcar filter with a width of 2.5 mm. We infer that the observed similarity of neighboring inactivation deficits (in Figure 6A) is approximately consistent with a combined effect of the known spatial spread of muscimol (~ 2.5 mm; Arikan et al., 2002) and the previously described phenomenon of anatomically neighboring neurons exhibiting similar patterns of responses (see Discussion).

Interestingly, although we tested 6–10 subtasks, the observed patterns of deficits could be reasonably well captured by a lower number of dimensions. In particular, using principal components analysis (PCA) on all deficit patterns measured under Experiment 1 (6 subtasks, n = 10, 7 inactivation sites in monkeys M and P, respectively), we found that the first three (out of six) principal components (PCs) captured more than 90% of the total variance in the deficit patterns. Note that 8 other sites measured under Experiment 2 in monkey P were not included here, as those sites were tested with different stimuli (textureless objects) and different subtasks. We represented this low-dimensional deficit subspace as a color space by mapping the first three PCs to RGB values. Figure 7A shows the embedding of each of the six subtasks tested in both monkeys in this color space, while the top panel of Figure 7B shows the deficit patterns of 17 inactivation sites projected onto this color space. The bottom panels of Figure 7B show the anatomical location of the same 17 inactivation sites overlaid on inflated cortical surfaces, each colored according to the corresponding RGB values obtained from this color space. Consistent with the inferred topographical organization in Figure 6A, we observe spatial clustering of colors in this map, suggesting that individual sites are involved in different combinations of subtasks, but that nearby sites are similarly involved (i.e., similarly colored).

Neurally Mechanistic Models That Link IT Activity to Behavior

Given the observed tissue specificity, we asked to what extent the observed behavioral deficits could be predicted by the neuronal activity patterns in the inactivated subregions (e.g.,

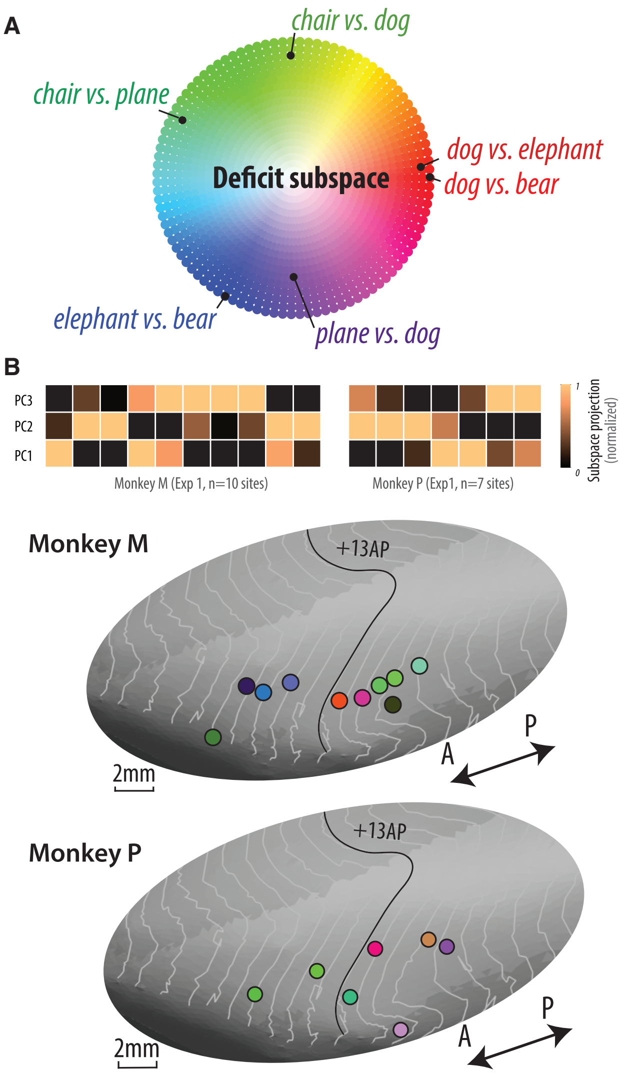

Figure 7. Visualization of Topographical Organization

Using dimensionality reduction on the behavioral deficit patterns across 17 inactivation sites and 6 subtasks, we obtained a three-dimensional deficit subspace that captured the majority of variance in deficit patterns. We represented this deficit subspace as a color space, by mapping the first three PCs to RGB values.

(A) The embedding of each of the six subtasks tested in both monkeys in this color space; the hue and saturation of colors are mapped to polar angle and eccentricity in this visualization, while the third dimension is not shown.

(B)The heatmap (top) shows the deficit patterns of 17 inactivation sites projected onto the deficit subspace (i.e., the first three PCs); for this visualization, subspace projections have been normalized to span $[0, 1]$ for each site. Each flattened map (bottom) shows the anatomical location of inactivation sites, each colored according to the corresponding RGB values obtained from this deficit subspace. Note that only sites measured under Experiment 1 (n = 10 in monkey M, n = 7 in monkey P) are shown; 8 other sites measured under Experiment 2 in monkey P were not included here, as those sites were tested with different stimuli (textureless objects). Consistent with the inferred topographical organization, we observe spatial clustering of colors in this map, suggesting that individual sites are involved in some combination of subtasks, and that neighboring sites are similarly involved (i.e., similarly colored).

prior to inactivation). The central panel in Figure 8A shows the location of an example muscimol inactivation site and local electrophysiology sites, co-registered using stereo micro-focal X-ray reconstruction, and overlaid on a coronal MRI slice. For this example site in IT, we recorded the activity of eight multi-unit sites (shown as cyan discs) in close proximity to the injection site (shown as red disc). Multi-unit activity was recorded in response to the same images as those used in behavioral testing, in a passive viewing paradigm (see STAR Methods). Each sub-panel shows a multi-unit site's stimulus-locked firing-rate responses for each of the five objects, averaged over images. We note that neuronal sites, while heterogeneous, each exhibit reliable object preferences. Based on local neuronal responses such as this, we constructed and tested a number of linking (a.k.a. "decoder") models, each of which maps the local IT spiking response patterns to a predicted behavioral deficit.

The right panel in Figure 8A shows the predictions from two example linking models. The local neural response models predict large deficits for subtasks with images that produce the largest response from the local neuronal sites. The local neural discriminability models predict large deficits for subtasks for which the local neural spiking activity was most discriminative, as measured by a linear classifier. We qualitatively observe that the discriminability models better capture the observed behavioral deficit patterns than the response models for this example inactivation site. This is quantified in Figure 8B as a noise-adjusted correlation between predicted and actual behavioral deficits over all inactivation sites with local neuronal recordings ($n = 11$ sites). All discriminability models significantly predict the inactivation deficits ($p < 0.0001$), while the response models failed to do so ($p > 0.05$). This result was consistent across the two animals (Figure S3C). For each inactivation site and decoding model, we tuned the distance threshold that was used to select neighboring neuronal sites to best predict the inactivation deficits; optimal distance thresholds ($\theta_d = 2.87 \pm 0.264mm$) approximately corresponded to the known spatial spread of muscimol in the cortical tissue (Arikan et al., 2002). Importantly, decoding models constructed from distant neural responses failed to significantly predict inactivation patterns ($p > 0.05$; see Figure 8B, dark bars). In summary, inactivation of millimeter-scale regions of IT results in behavioral deficits that are predicted by the local neuronal discriminability.

DISCUSSION

In this work, we sought to investigate *whether* and *how* neural activity in IT causally supports core object recognition behavior. Specifically, our goals were to (1) directly test the hypothesis that IT is a necessary node in the brain's neural network that underlies potentially all core object recognition discrimination behavior (subtasks), and (2) to ask whether any such causal role is functionally organized over the cortical tissue. To this end, we reversibly inactivated individual, arbitrarily sampled millimeter-scale regions of IT while monkeys performed a battery of arbitrarily sampled basic-level object discrimination subtasks. With the explicit goal of testing "arbitrary" basic-level object recognition, these subtasks were pseudo-randomly selected from the large set of pairwise discriminations that these animals were

Please cite this article in press as: Rajalingham and DiCarlo, Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits, Neuron (2019), https://doi.org/10.1016/j.neuron.2019.02.001

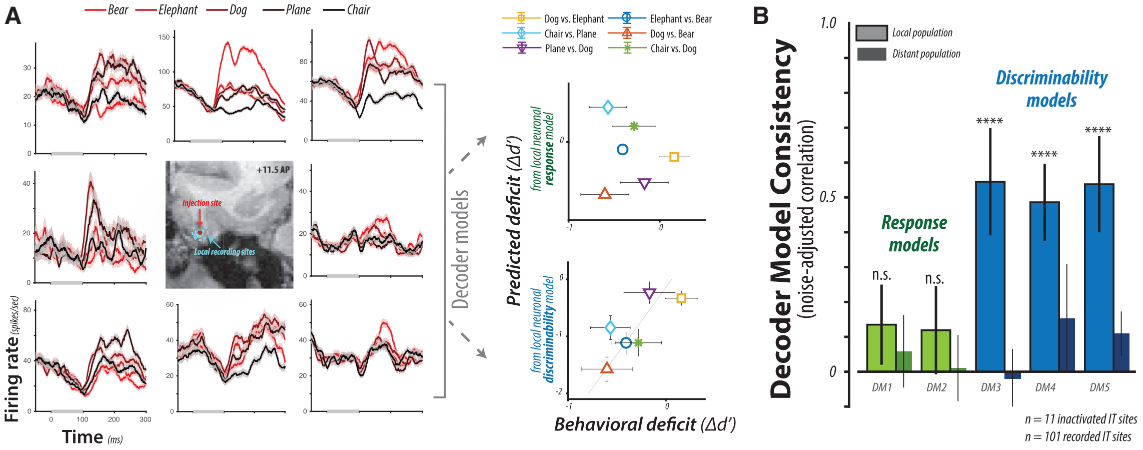

Figure 8. Relationship of IT Spiking Responses to Patterns of Behavioral Deficits

(A) Left: multi-unit spiking activity recorded serially (prior to inactivation) with a single microelectrode for eight sites sampled within an example IT subregion. Center: the recording locations (each determined via stereo, micro-focal X-ray; see STAR Methods) are plotted here projected into the plane of a single MRI slice containing the center of the IT inactivated region. Neural responses were measured in a rapid-serial-visual-presentation (RSVP) paradigm with 100 ms on and 100 ms off. Responses were averaged across all images and all repetitions for each object. The decreasing response prior to stimulus onset simply reflects the offset of the stimulus presented immediately before. Each inset panel shows the spiking activity response to each of five objects aligned to stimulus onset; each line is the mean activity averaged over all images of each object and all repetitions (40 images/object, ~ 10 repetitions/image), and the shaded region corresponds to SEM. Gray bar shows image presentation time (100 ms). Neuronal sites, while heterogeneous, each exhibit object preferences, even when averaging over images. Right: to determine whether the observed behavioral deficits are predicted by local neuronal activity, we constructed and tested several decoder models that transform IT response patterns from these 8 multi-unit sites into predictions of behavioral deficits resulting from inactivation (see STAR Methods). The predictions of two of these models (upper and lower scatterplot) are compared with the measured behavioral deficits for this example IT-inactivation site. Note that larger deficits correspond to more negative values of Δd' (lower left corner of each scatterplot).

(B) The average predictive power of each of five tested decoder model is shown as the noise-adjusted correlation between predicted and actual behavioral deficits for all relevant sites (i.e., where we had both the local spiking responses [as in A] and the pattern of behavioral deficits measure on the same set of images). Each light-colored bar corresponds to a specific decoding model (models DM1–DM5, see STAR Methods) constructed from the local neuronal population. All local neuronal discriminability models (blue) were clearly better than the local neuronal response models (green). Dark-colored bars correspond to the exact same decoding models, constructed from the most-distant neuronal population (i.e., the neuronal population of the most anatomically distant inactivation site). All decoding models constructed from distant neural population failed to significantly predict inactivation patterns ($p > 0.05$).

previously trained on and did not include any face-related discriminations. Our first contribution is to provide new direct causal evidence for the role of IT in core object recognition, which was both scarce and equivocal, especially beyond the specific case of face-selective subregions of IT. Moreover, our results revealed that the causal role of IT in object recognition has topographic organization at the millimeter scale and is predicted by local neuronal discriminability. With respect to the outline in Figure 1A, our data are sufficient to distinguish between the alternative outcomes (despite not directly testing face-related discriminations), and strongly support the heterogeneous deficit pattern (rightmost panel in Figure 1A). Together, these advances solidify the previously presumed causal role of the IT cortex in core object recognition and could be used to distinguish among alternative neurally mechanistic (i.e., neural network) models of the ventral stream and its role in core object recognition behavior, as outlined below.

The Hypothesized Role of the IT Cortex in Core Object Recognition Behavior

Here, we define the decoding hypothesis (a.k.a. linking hypothesis; Brindley, 1960) that motivated the present study and alternatives to that hypothesis. First, we hypothesize that the IT cortex is a necessary node in the brain's neural network that underlies core recognition behavior (Prediction 1). Stated in other words, our hypothesis is that core object recognition behavior causally depends on the firing of neurons in the IT cortex, and, without those spikes, core object recognition behavior would be at chance (DiCarlo et al., 2012; Majaj et al., 2015). Importantly, core object recognition behavior is not a single subtask, but is a domain of many possible subtasks, including at least hundreds of pairwise object discrimination subtasks in monkeys (Rajalingham et al., 2015, 2018). Thus, based on prior IT recording work (Majaj et al., 2015), our decoding hypothesis is more specific: each IT neuron is a necessary part of multiple such subtasks (Prediction 2), which is contrasted with the alternative possibility that all non-face-selective IT neurons are necessary for *all* non-face-related subtasks. Third, our decoding hypothesis is that single IT neurons that carry information that might *potentially* support each subtask are indeed *necessary* for each such subtask, and they are necessary *regardless of their physical location in IT* (Prediction 3). This hypothesis is implicitly stated in (Majaj et al., 2015) and explicitly discussed in (Afraz et al., 2015). However, because prior work (Tanaka, 1996; Kreiman et al., 2006; Sato et al., 2009) showed that IT neurons with similar object feature and image preferences tend to be clustered at millimeter scale, our decoding hypothesis (above) predicts that each millimeter-scale IT subregion is an enrichment of neurons that are necessary nodes in some object discrimination subtasks (again, more than one subtask). This is contrasted with the alternative possibility that each subregion of IT is equally involved in all object discrimination subtasks. We note that all of these assumptions (here, collectively called our ''decoding hypothesis'') and the resultant predictions (Predictions 1–3) were in place prior to our undertaking of this study, and, indeed, were the motivation of this study.

Direct Causal Evidence for the Role of IT in Core Object Recognition

While we cannot yet test all the mechanistic aspects of this decoding hypothesis (above), we can test some of its most basic predictions; to our knowledge, these tests had not yet been done. To carry out these tests, we adopt the terminology of Jazayeri and Afraz (2017), whereby ''causal'' dependencies can be inferred by correlating a dependent variable to an experimentally controlled variable, in contrast to correlational dependencies, which are associations between variables that we measure and may indirectly control, but we do not directly control. Thus, to infer a causal link between IT activity and behavior, it is necessary to specifically manipulate activity in IT (e.g., via the application of pharmacological agents into IT to silence neurons, etc.) while measuring behavior. Related correlational dependencies (e.g., via direct manipulation of visual input to the retinae while measuring variations from both IT activity and behavior) are consistent with our causal decoding hypothesis (outlined above) but could also be epiphenomenal (i.e., the resulting IT activity caused by the stimulus is correlated with, but does not cause, the behavior). Recently, research in other behavioral domains has exposed divergences between correlational and causal dependencies (Katz et al., 2016; Liu and Pack, 2017), highlighting the need to directly test causal dependencies.

With respect to Prediction 1 of our stated decoding hypothesis (that IT is necessary for core object recognition), decades of neurophysiological and neuropsychological research suggest that activity in IT cortex is a good neural correlate of primate object recognition behavior (Logothetis and Sheinberg, 1996; Tanaka, 1996; Rolls, 2000; DiCarlo et al., 2012): individual neurons in the IT cortex are selective to complex visual features in images and exhibit remarkable tolerance to changes in viewing parameters (Kobatake and Tanaka, 1994; Ito et al., 1995; Logothetis et al., 1995; Booth and Rolls, 1998; Rust and DiCarlo, 2010), and the population of neurons in IT not only matches overall primate behavioral performance (Hung et al., 2005; Zhang et al., 2011) but also reliably predicts the behavioral performance on each subtask (Majaj et al., 2015). Taken together, these results are consistent with our decoding hypothesis, but could also be epiphenomenal. To this end, our first major contribution in this work is to provide direct evidence in support of Prediction 1.

Prior to this, causal evidence for the role of IT in core object recognition has been both scarce and equivocal, especially beyond the specific case of face-selective regions in IT. Lesions of IT suggest a coarse causal link between this area and visual behaviors (Cowey and Gross, 1970; Manning, 1972; Holmes and Gross, 1984; Weiskrantz and Saunders, 1984; Buffalo et al., 1998; Huxlin et al., 2010; Matsumoto et al., 2016), but the resulting behavioral deficits are often contradictory (Dean, 1974; Huxlin et al., 2010) and at best modest (Horel et al., 1987; Matsumoto et al., 2016). For example, recent work showed that near complete ablation of IT (bilateral removal of anterior IT) resulted in only mild (10%–15%) deficits in object categorization (Matsumoto et al., 2016). Such modest behavioral deficits from large-scale ablations may be due to limitations of the methodologies and the behavioral assays, both of which may not be robust to compensatory neural mechanisms. For example, other visual cortical areas could be recruited via post-lesion neural plasticity, and behavioral tasks that explicitly require viewpoint invariance may help mitigate such concerns (Weiskrantz and Saunders, 1984). In this work, we did not test a directly comparable experimental paradigm (e.g., via complete pharmacological inactivations of all of IT cortex). Rather, we made focal inactivations of IT subregions and observed deficits whose effect sizes were largely commensurate with the very small size of these inactivations. A handful of studies have reported using focal reversible neural perturbation tools (e.g., electrical, pharmacological, and optogenetic perturbation) to test the stated decoding hypothesis, but all exclusively targeted spatial clusters of face-selective neurons in IT, testing the causal role of these regions in basic- and subordinate-level face recognition behaviors (Afraz et al., 2006, 2015; Moeller et al., 2017; Sadagopan et al., 2017), with one notable exception (Verhoef et al., 2012). Thus, our results provide the most systematic direct causal evidence for the general decoding hypothesis (i.e., Prediction 1) outlined above.

The Causal Role of IT in Core Object Recognition Is Topographically Organized

With respect to Prediction 2 of our stated decoding hypothesis (that each millimeter-scale IT subregion is necessary for several, but not all, object discrimination subtasks), our second major contribution in this work is to provide direct evidence for a subtask-selective causal role of IT in core object recognition at the millimeter-scale. Prior to this, all existing studies have exclusively targeted specific spatial clusters of face-selective neurons in IT, testing the causal role of these regions in basic- and subordinate-level face recognition behaviors (Afraz et al., 2006, 2015; Moeller et al., 2017; Sadagopan et al., 2017; Verhoef et al., 2012). While faces are an especially behaviorally relevant stimulus class for primates (Tsao and Livingstone, 2008), the experimental bias toward spatial clusters in IT may also reflect the spatial resolution limitations of current neural perturbation tools, which operate on groups of spatially contiguous neurons at approximately millimeter-scale. Given this limitation, the known millimeter-scale spatial clusters of face-selective regions in IT (Tsao et al., 2003, 2006; Tsao and Livingstone, 2008) form an intuitively optimal candidate for testing causal dependencies related to our decoding hypothesis. We note that similar spatial clustering of response selectivity has been reported for a small number of other image groupings besides faces, such as color, disparity, places, and bodies (Conway et al., 2007; Kornblith et al., 2013; Lafer-Sousa and Conway, 2013; Verhoef et al., 2015; Popivanov et al., 2012).

Importantly, the topographic organization of neurons in IT is largely unknown and assumed by many to be functionally random and non-specific beyond these discrete clusters. To support a general inference, we tested arbitrary sampled millimeter-scale regions of ventral IT rather than functionally target inactivation sites. This highlights an important and novel contribution of our work in testing multiple regions on multiple subtasks and making inferences into the organization of such subtasks over the cortical tissue. Interestingly, we found that inactivation of different regions in ventral IT led to different subtask-specific deficits, suggesting some functional specificity for arbitrarily sampled millimeter-scale regions. Indeed, our data suggest that each millimeter-scale region in IT is causally involved in a relatively small proportion ($\sim$ 25%) of object recognition subtasks, and that anatomically neighboring regions are similar in this regard. Given that we targeted all inactivations to mm-scale regions of the ventral surface of IT, it is possible that other regions in IT (e.g., in the STS or on the lateral surface) could in principle result in much larger or more selective deficits on these subtasks. However, there is currently no evidence for or against this claim, primarily due to a dearth of comparable inactivation studies. (Note that Afraz et al. [2015] and Sadagopan et al. [2017] differ with regards to stimulus lateralization and injection volume but are not inconsistent with our findings.) The causal topographical organization inferred from our results is consistent with previously reported sub-millimeter scale columnar organization of neurons in IT (Fujita et al., 1992; Tanaka, 1996; Wang et al., 1996, 1998; Kreiman et al., 2006) and broader eccentricity-dependent organization of IT cortex (Conway, 2018). We speculate that this topographic organization could reflect a general principle of global cortical layout, whereby neuronal selectivities are developed in the face of metabolic constraints (e.g., minimization of connection wiring length; Chklovskii et al., 2002).

The causal role of IT in core object recognition is predicted by the local neuronal discriminability. Finally, with respect to Prediction 3 of our decoding hypothesis, we found that behavioral deficits from inactivating millimeter-scale regions of IT are consistent with predictions from a spatially distributed readout of neurons in IT (Majaj et al., 2015). Indeed, inactivation deficits were well predicted by local neuronal discriminability decoding models, suggesting that the causal role of each IT subregion is well approximated by the information that is coded explicitly (i.e., linearly separable) by the local population of neurons. In contrast, inactivation deficits were not well predicted by specific local neural response readout models, which predict that neurons that respond highly to particular stimulus classes, without encoding the differences between them in a linearly separable manner, are causally involved in discrimination between these classes. None of the tested decoding models perfectly explain the inactivation deficits, potentially due to data limits. In the current work, we did not have sufficient neural sampling to directly test population decoding models (e.g., by simulating perturbations on a localized sub-population within a representative sample of all of IT and measuring the resulting simulated behavior). Nevertheless, our results are consistent with at least one decoding hypothesis (Majaj et al., 2015) (Figure 8B).

Importantly, our results speak directly to questions of long-standing interest in systems and cognitive neuroscience, in particular for the human neuroimaging community. A belief held by many in this field is that the overall responsiveness of a cluster of neurons is indicative of its causal role in behavior. For example, one might conclude that face-selective regions, which respond preferentially to images of faces, must causally support face detection and discrimination behaviors (Tsao and Livingstone, 2008). Similarly, one might make the inverse inference that clusters of neurons that don't preferentially respond to faces on average, regardless of whether they contain explicit subtask-relevant information, do not causally support such face-related subtasks. An alternative hypothesis, used as the basis for techniques such as multi-voxel pattern analysis (MVPA) (Norman et al., 2006), is that the behavioral role of a cluster of neurons is not determined by its responsiveness, per se, but by its discriminability (i.e., its subtask-explicit information content). However, it is still under debate whether such explicitly available information carried by that cluster of neurons is in fact "used" by the brain to produce behavior, or is epiphenomenal (Williams et al., 2007).

To date, there have been a handful of attempts in human cognitive neuroscience to resolve this debate and discriminate between these alternative hypotheses using coarse perturbations of neural activity (e.g., transcranial magnetic stimulation and electrical stimulation; Parvizi et al., 2012; Schalk et al., 2017; Pitcher et al., 2007). In this study, we were able to both record from and inactivate the neuronal activity in arbitrarily selected millimeter-scale regions in primate IT. In contrast to previous human cognitive neuroscience studies, we found that responsiveness is not at all predictive of the behavioral deficits resulting from inactivation. Instead, our results are consistent with a decoding hypothesis based on neuronal discriminability (Majaj et al., 2015; Afraz et al., 2015) and demonstrate that one should not conclude that a cluster of neurons that preferentially responds to a particular group of stimuli causally supports the ability to discriminate between stimuli within that group.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Subjects and surgery
- METHOD DETAILS
  - ○ Core object recognition behavioral paradigm
  - ○ Test images
  - ○ Physiology and pharmacology
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Behavioral metrics
  - ○ Sparsity of deficit
  - ○ Neuronal readout models
  - ○ Noise-adjusted correlations
  - ○ Statistical testing
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found with this article online at https://doi.org/10.1016/j.neuron.2019.02.001.

ACKNOWLEDGMENTS

We thank Arash Afraz, Nancy Kanwisher, and Roozbeh Kiani for useful conversations, and Elias Issa for access to previously published data. This research was supported by US National Eye Institute grants R01-EY014970 (J.J.D.), Office of Naval Research MURI-114407 (J.J.D), and the Simons Foundation SCGB (325500) (J.J.D.).

AUTHOR CONTRIBUTIONS

R.R. and J.J.D. designed research; R.R. performed research and analyzed data; and R.R. and J.J.D. wrote the paper.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 11, 2018
Revised: December 6, 2018
Accepted: January 31, 2019
Published: March 13, 2019

SUPPORTING CITATIONS

The following references appear in the Supplemental Information: Issa et al., 2018.

REFERENCES

Afraz, A., Boyden, E.S., and DiCarlo, J.J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. Proc. Natl. Acad. Sci. USA *112*, 6730–6735.

Afraz, S.-R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. Nature *442*, 692–695.

Andrews, P.R., and Johnston, G.A.R. (1979). GABA agonists and antagonists. Biochem. Pharmacol. *28*, 2697–2702.

Arikan, R., Blake, N.M.J., Erinjeri, J.P., Woolsey, T.A., Giraud, L., and Highstein, S.M. (2002). A method to measure the effective spread of focally injected muscimol into the central nervous system with electrophysiology and light microscopy. J. Neurosci. Methods *118*, 51–57.

Biederman, I., Gerhardstein, P.C., Cooper, E.E., and Nelson, C.A. (1997). High level object recognition without an anterior inferior temporal lobe. Neuropsychologia *35*, 271–287.

Booth, M.C., and Rolls, E.T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. Cereb. Cortex *8*, 510–523.

Brindley, G.S. (1960). Physiology of the Retina and the Visual Pathway (Williams & Wakins).

Buffalo, E.A., Stefanacci, L., Squire, L.R., and Zola, S.M. (1998). A reexamination of the concurrent discrimination learning task: the importance of anterior inferotemporal cortex, area te. Behav. Neurosci. *112*, 3.

Buffalo, E.A., Ramus, S.J., Squire, L.R., and Zola, S.M. (2000). Perception and recognition memory in monkeys following lesions of area te and perirhinal cortex. Learn. Mem. *7*, 375–382.

Chklovskii, D.B., Schikorski, T., and Stevens, C.F. (2002). Wiring optimization in cortical circuits. Neuron *34*, 341–347.

Conway, B.R. (2018). The organization and operation of inferior temporal cortex. Annual review of vision science *4*, 381–402.

Conway, B.R., Moeller, S., and Tsao, D.Y. (2007). Specialized color modules in macaque extrastriate cortex. Neuron *56*, 560–573.

Cowey, A., and Gross, C.G. (1970). Effects of foveal prestriate and inferotemporal lesions on visual discrimination by rhesus monkeys. Exp. Brain Res. *11*, 128–144.

Cox, D.D., Papanastassiou, A.M., Oreper, D., Andken, B.B., and DiCarlo, J.J. (2008). High-resolution three-dimensional microelectrode brain mapping using stereo microfocal x-ray imaging. J. Neurophysiol. *100*, 2966–2976.

Dean, P. (1974). The effect of inferotemporal lesions on memory for visual stimuli in rhesus monkeys. Brain Res. *77*, 451–469.

DiCarlo, J.J., and Johnson, K.O. (1999). Velocity invariance of receptive field structure in somatosensory cortical area 3b of the alert monkey. J. Neurosci. *19*, 401–419.

DiCarlo, J.J., and Cox, D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341.

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

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. Nature *360*, 343–346.

Green, D.M., and Swets, J.A. (1966). Signal detection theory and psychophysics*Volume 1* (Wiley).

Holmes, E.J., and Gross, C.G. (1984). Effects of inferior temporal lesions on discrimination of stimuli differing in orientation. J. Neurosci. *4*, 3063–3068.

Horel, J.A., Pytko-Joiner, D.E., Voytko, M.L., and Salsbury, K. (1987). The performance of visual tasks while segments of the inferotemporal cortex are suppressed by cold. Behav. Brain Res. *23*, 29–42.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866.

Huxlin, K.R., Saunders, R.C., Marchionini, D., Pham, H.-A., and Merigan, W.H. (2010). Perceptual deficits after lesions of inferotemporal cortex in macaques. Cereb. Cortex *10*, 671–683.

Issa, E.B., Papanastassiou, A.M., and DiCarlo, J.J. (2013a). Large-scale, high-resolution neurophysiological maps underlying FMRI of macaque temporal lobe. J. Neurosci. *33*, 15207–15219.

Issa, E.B., Papanastassiou, A.M., and DiCarlo, J.J. (2013b). Large-scale, high-resolution neurophysiological maps underlying FMRI of macaque temporal lobe. J. Neurosci. *33*, 15207–15219.

Issa, E.B., Cadieu, C.F., and DiCarlo, J.J. (2018). Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. bioRxiv.

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. J. Neurophysiol. *73*, 218–226.

Jazayeri, M., and Afraz, A. (2017). Navigating the neural space in search of the neural code. Neuron *93*, 1003–1014.

Johnson, K.O., Hsiao, S.S., and Yoshioka, T. (2002). Neural coding and the basic law of psychophysics. Neuroscientist *8*, 111–121.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. USA *107*, 11163–11170.

Katz, L.N., Yates, J.L., Pillow, J.W., and Huk, A.C. (2016). Dissociated functional significance of decision-related activity in the primate dorsal stream. Nature *535*, 285–288.

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

Kornblith, S., Cheng, X., Ohayon, S., and Tsao, D.Y. (2013). A network for scene processing in the macaque temporal lobe. Neuron *79*, 766–781.

Kreiman, G., Hung, C.P., Kraskov, A., Quiroga, R.Q., Poggio, T., and DiCarlo, J.J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. Neuron *49*, 433–445.

Lafer-Sousa, R., and Conway, B.R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. Nat. Neurosci. *16*, 1870–1878.

Liu, L.D., and Pack, C.C. (2017). The contribution of area mt to visual motion perception depends on training. Neuron *95*, 436–446.e3.

Logothetis, N.K., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. Curr. Biol. *5*, 552–563.

Logothetis, N.K., and Sheinberg, D.L. (1996). Visual object recognition. Annu. Rev. Neurosci. *19*, 577–621.

Macmillan, N.A. (1993). Signal detection theory as data analysis method and psychological decision model. In A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues, G. Keren and C. Lewis, eds. (Lawrence Erlbaum Associates, Inc), pp. 21–57.

Majaj, N.J., Hong, H., Solomon, E.A., and DiCarlo, J.J. (2015). Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. J. Neurosci. *35*, 13402–13418.

Manning, F.J. (1972). Serial reversal learning by monkeys with inferotemporal or foveal prestriate lesions. Physiol. Behav. *8*, 177–181.

Matsumoto, N., Eldridge, M.A.G., Saunders, R.C., Reoli, R., and Richmond, B.J. (2016). Mild perceptual categorization deficits follow bilateral removal of anterior inferior temporal cortex in rhesus monkeys. J. Neurosci. *36*, 43–53.

Moeller, S., Crapse, T., Chang, L., and Tsao, D.Y. (2017). The effect of face patch microstimulation on perception of faces and objects. Nature Neuroscience *20*, 743–752.

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. *10*, 424–430.

Noudoost, B., and Moore, T. (2011). A reliable microinjectrode system for use in behaving monkeys. J. Neurosci. Methods *194*, 218–223.

Op de Beeck, H., and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. J. Comp. Neurol. *426*, 505–518.

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

Parvizi, J., Jacques, C., Foster, B.L., Witthoft, N., Rangarajan, V., Weiner, K.S., and Grill-Spector, K. (2012). Electrical stimulation of human fusiform face-selective regions distorts face perception. J. Neurosci. *32*, 14915–14920.

Pinto, N., Cox, D.D., and DiCarlo, J.J. (2008). Why is real-world visual object recognition hard? PLoS Comput. Biol. *4*, e27.

Pitcher, D., Walsh, V., Yovel, G., and Duchaine, B. (2007). TMS evidence for the involvement of the right occipital face area in early face processing. Curr. Biol. *17*, 1568–1573.

Popivanov, I.D., Jastorff, J., Vanduffel, W., and Vogels, R. (2012). Stimulus representations in body-selective regions of the macaque cortex assessed with event-related fMRI. Neuroimage *63*, 723–741.

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

Rajalingham, R., Issa, E.B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J.J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J. Neurosci. *38*, 7255–7269.

Rolls, E.T. (2000). Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. Neuron *27*, 205–218.

Rust, N.C., and DiCarlo, J.J. (2010). Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area v4 to it. J. Neurosci. *30*, 12978–12995.

Sadagopan, S., Zarco, W., and Freiwald, W.A. (2017). A causal relationship between face-patch activity and face-detection behavior. eLife *6*, e18558.

Sato, T., Uchida, G., and Tanifuji, M. (2009). Cortical columnar organization is reconsidered in inferior temporal cortex. Cereb. Cortex *19*, 1870–1888.

Schalk, G., Kapeller, C., Guger, C., Ogawa, H., Hiroshima, S., Lafer-Sousa, R., Saygin, Z.M., Kamada, K., and Kanwisher, N. (2017). Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. Proc. Natl. Acad. Sci. USA *114*, 12285–12290.

Sheinberg, D.L., and Logothetis, N.K. (1997). The role of temporal cortical areas in perceptual organization. Proc. Natl. Acad. Sci. USA *94*, 3408–3413.

Tanaka, K. (1996). Inferotemporal cortex and object vision. Annu. Rev. Neurosci. *19*, 109–139.

Tsao, D.Y., Freiwald, W.A., Knutsen, T.A., Mandeville, J.B., and Tootell, R.B. (2003). Faces and objects in macaque cerebral cortex. Nat. Neurosci. *6*, 989–995.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670–674.

Tsao, D.Y., and Livingstone, M.S. (2008). Mechanisms of face perception. Annu. Rev. Neurosci. *31*, 411–437.

Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. Nat. Neurosci. *4*, 832–838.

Ullman, S. (1996). High-level vision: Object Recognition and Visual Cognition*Volume 2* (MIT Press).

Verhoef, B.-E., Vogels, R., and Janssen, P. (2012). Inferotemporal cortex subserves three-dimensional structure categorization. Neuron *73*, 171–182.

Verhoef, B.E., Bohon, K.S., and Conway, B.R. (2015). Functional architecture for disparity in macaque inferior temporal cortex and its relationship to the architecture for faces, color, scenes, and visual field. J. Neurosci. *35*, 6952–6968.

Vinje, W.E., and Gallant, J.L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. Science *287*, 1273–1276.

Wang, G., Tanaka, K., and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex. Science *14*, 272.

Wang, G., Tanifuji, M., and Tanaka, K. (1998). Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. Neurosci. Res. *32*, 33–46.

Weiskrantz, L., and Saunders, R.C. (1984). Impairments of visual object transforms in monkeys. Brain *107*, 1033–1072.

Williams, M.A., Dang, S., and Kanwisher, N.G. (2007). Only some spatial patterns of fMRI response are read out in task performance. Nat. Neurosci. *10*, 685–686.

Zhang, Y., Meyers, E.M., Bichot, N.P., Serre, T., Poggio, T.A., and Desimone, R. (2011). Object decoding with attention in inferior temporal cortex. Proc. Natl. Acad. Sci. USA *108*, 8850–8855.

Please cite this article in press as: Rajalingham and DiCarlo, Reversible Inactivation of Different Millimeter-Scale Regions of Primate IT Results in Different Patterns of Core Object Recognition Deficits, Neuron (2019), https://doi.org/10.1016/j.neuron.2019.02.001

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Organisms/Strains		
Rhesus monkeys	California National Primate Research Center	https://cnprc.ucdavis.edu/
Software and Algorithms		
MATLAB	MathWorks	MATLAB 9.2 (R2017a)

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, James J. DiCarlo (dicarlo@mit.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Subjects and surgery

Two adult male rhesus macaque monkeys (Macaca mulatta, subjects M, P) were trained on the core object recognition paradigm described below. For each animal, a surgery using sterile technique was performed under general anesthesia to implant a titanium head post to the skull using titanium screws, and a cylindrical recording chamber (19 mm inner diameter; Crist Instruments) over a craniotomy targeting the temporal lobe in the left hemisphere from the top of the skull (Monkey M, +13 mm posterior-anterior, +16.3 mm medial-lateral, 15° medial-lateral angle; Monkey P, +13 mm posterior-anterior, +14.75 mm medial-lateral, 15° medial-lateral angle). All procedures were performed in compliance with the guideline of National Institutes of Health and the American Physiological Society, and approved by the MIT Committee on Animal Care.

METHOD DETAILS

Core object recognition behavioral paradigm

Core object discrimination is defined as the ability to discriminate between two or more objects in visual images presented under high view uncertainty in the central visual field (~ 10°), for durations that approximate the typical primate, free-viewing fixation duration (~ 200 ms) (DiCarlo and Cox, 2007; DiCarlo et al., 2012). As in our previous work (Rajalingham et al., 2015, 2018), we investigate this behavior using batteries of trial-by-trial interleaved set of pairwise object discrimination subtasks. The behavioral paradigm is described below. Behavioral data was collected under head fixation, and subjects reported their choices using their gaze. We monitored eye position by tracking the position of the pupil using a camera-based system (SR Research Eyelink 1000). Images were presented on a 27'' LCD monitor (1920 × 1080 at 60 Hz; Samsung S27A850D) positioned 44 cm in front of the animal. At the start of each training session, subjects performed an eye-tracking calibration subtask by saccading to a range of spatial targets and maintaining fixation for 800ms. Calibration was repeated if drift was noticed over the course of the session.

Figure 1B illustrates the behavioral paradigm. Each trial was initiated when the monkey acquired and held gaze fixation on a central fixation point for 200ms, after which a test image (8×8° of visual angle in size) appeared at the center of gaze for 100ms. Trials were aborted if gaze was not held within ±2°. After extinction of the test image, two choice images, each displaying a single object in a canonical view with no background, were immediately shown to the left and right (each centered at 8° of eccentricity along the horizontal meridian; see Figure 1B). One of these two objects was always the same as the object that generated the test image (i.e., the correct choice), and its location (left or right) was randomly chosen on each trial. The object that was not displayed in the test image is referred to as the distractor object, but note that objects are equally likely to be distractors and targets. The monkey was allowed to freely view the choice images for up to 1000ms, and indicated its final choice by holding fixation over the selected image for 700ms. The monkey was rewarded with a small juice reward for each correct trial. After the end of each trial, another fixation point appeared, cueing the next trial. Each trial consisted of a different randomly selected pairwise object discrimination subtask. Note that each pairwise subtask is operationally defined by the pair of choice objects at the end of the trial, and we ensure that the test images are chosen in a balanced way such that approximately half of the trials begin with test images of one object and the other half of the trials begin with test images of the other object. Performance of each such "pairwise subtask" is the primary unit of measure in this study (averaged over all test images of each object, unless otherwise noted). Note that, because the trials of each such pairwise discrimination subtask are randomly interleaved, the subject cannot anticipate which object will be shown or which pair of object choices will appear after the test image. Real-time experiments for monkey psychophysics were controlled by open-source software (MWorks Project http://mworks-project.org/).

Both animals were previously trained on other images of other objects, and were proficient in discriminating among over 35 arbitrarily sampled basic-level object categories (i.e., several hundreds of possible pairwise object discrimination subtasks). In this study, five randomly selected basic-level objects were tested (bear, elephant, dog, airplane, and chair). While other (unpublished) work suggests that more objects are needed to fully exercise the domain of core object recognition, in this study our primary goal was to balance between spanning that domain and collecting enough behavioral trials to detect even subtle changes in discrimination performance that might result from suppression of IT subregions. Our choice of five objects resulted in ten possible pairwise object discrimination subtasks (see Figure 1A for complete list). To accumulate enough trials to precisely measure performance for each subtask within a single behavioral session (i.e., a single experimental day), we sub-selected six of these ten subtasks for most experiments. For a subset of experiments in one animal (monkey P, experiment 2), we tested all ten pairwise subtasks. For each session, monkeys were tested for several hours (until satiation) and performed a large number of trials (monkey M: 3442 ± 1097, monkey P: 4430 ± 942; mean ± SD).

Test images

We examined basic-level object recognition behavior by generating test images of the five objects (above) that were synthesized from the five computer models of each object. As in prior work (Rajalingham et al., 2015, Majaj et al., 2015), the goal was to use naturalistic images that also exercised the view invariance challenges of core object recognition, as such images are excellent at differentiating between low-level representations and primate behavior. The image generation pipeline is described in detail elsewhere (Majaj et al., 2015). Briefly, each image was generated by first rendering the object with randomly chosen viewing parameters (2D position, 3D rotation and viewing distance), and then placing that foreground object view onto a randomly chosen, natural image. Object models spanned basic-level object categories (bear, elephant, dog, airplane, and chair). Background images were sampled randomly from a large database of high-dynamic range images of indoor and outdoor scenes obtained from Dosch Design (http://www.doschdesign.com). This image generation procedure enforces invariant object recognition as it requires the animal to tackle the invariance problem, the computational crux of object recognition (Ullman, 1996; Pinto et al., 2008). Note that this design is in contrast to many prior perturbation studies of IT cortex in which the subject is required only to match one *image* to that same image (a.k.a. standard ''match-to-sample'') (Horel et al., 1987; Biederman et al., 1997), while here the subject must match *any possible image* of an object to a visual token (canonical view) that stands for that object.

The majority of the behavioral data presented here were collected in response to a base image set generated from the five objects (40 test images of each object, 200 test images in total). We additionally generated a variant of this dataset consisting of texture-less images of the same objects. These texture-less images were targeted to both titrate the subtask difficulty and further remove potential low-level confounds (e.g., luminance and contrast). This texture-less image set was not held fixed in size: on each behavioral session, we tested subjects on a mixture of 20% previously seen and 80% completely novel texture-less images of the same five objects, to mitigate potential memorization strategies. For the purpose of the current work, we treat both of these image sets as equivalent, namely as images of the same five objects under study differing only in their precise generative parameters. Figure S1A shows example two images for each object, from both image sets.

Physiology and pharmacology

In each animal, we first recorded multi-unit activity (MUA) from randomly sampled sites on the ventral surface of IT (monkey M: 57 multi-unit sites, monkey P: 43 multi-unit sites). Recordings in each animal were made over a period of several weeks using glass-coated tungsten micro-electrodes (impedance, $0.3 - 0.5M\Omega$; outer diameter, 310um; Alpha Omega). A motorized micro-drive (Alpha Omega) was used to lower electrodes through a 26-gauge stainless-steel guide tube inserted into the brain (5 mm) and held by a plastic grid inside the recording chamber (CRIST). We recorded MUA responses from IT while monkeys passively fixated images in a rapid serial visual presentation (RSVP) protocol (10 images/trial, 100ms on, 100 ms off). To ensure accurate stimulus presentation, eye position was tracked and trials were aborted if gaze was not held within $\pm 1.5^{\circ}$. To ensure accurate stimulus locking, spikes were aligned to a photodiode trigger attached to the display screen. Multi-unit responses were amplified (1x head-stage), filtered (250Hz cutoff), digitized (sampling rate of 40kHz) and sorted (Plexon MAP system, Plexon Inc.). For each image and multi-unit site, the image response patterns were obtained by first averaging MUA over many ($\sim$ 10) image repetitions, and computing the number of repetition-averaged spikes in two post-stimulus windows (70-170ms, 170-270ms)

Following this mapping stage, we performed inactivation experiments using focal microinjections of muscimol, a potent GABA agonist (Andrews and Johnston, 1979). We varied the location of microinjections to randomly sample the ventral surface of IT (from approximately $+$ 8mm AP to approx $+$ 20mm AP). Given the relatively long half-life of muscimol, inactivation sessions were interleaved over days with control behavioral sessions. Thus, each inactivation *experiment* consisted of three behavioral sessions: the baseline or pre-control session (1 day prior to injection), the inactivation session, and the recovery or post-control session (2 days after injection). Each inactivation session began with a single focal microinjection of 1μl of muscimol (5mg/mL, Sigma Aldrich) at a slow rate (100nl/min) via a 30-gauge stainless-steel cannula at the targeted site in ventral IT. Injections were made through a simple microinjection circuit consisting of a three-way valve (Labsmith) and marker line (similar to [Noudoost and Moore, 2011]), enabling precise monitoring of the flow and volume of muscimol injected. In pilot experiments, we verified complete neural suppression at the location of injection using custom-built single-use injectrodes (Noudoost and Moore, 2011). Given this volume of muscimol, we estimate strong neural suppression within a local region of $\sim$ 2.5mm in diameter for up to six hours after injection

(Arikan et al., 2002). After completion of the injection, we waited 10-20 min before measuring the monkey's behavior on a battery of object recognition subtasks for up to 3 hours post-injection.

To ensure accurate targeting of IT and reconstruction of the relative positions of injection and recording locations, all electrophysiological recordings and pharmacological injections were made under micro-focal stereo X-ray guidance (Cox et al., 2008). Briefly, monkeys were fitted with a plastic frame (3 × 4 cm) positioned near the temporal lobe using a plastic arm anchored in the dental acrylic implant. The frame contained six brass fiducial markers (1mm diameter) of known geometry, measured using micro-CT. The fiducial markers formed a fixed 3D skull-based coordinate system for registering all physiological recordings and pharmacological injection sites. At each site, two X-rays were taken simultaneously at near orthogonal angles, and the 3D location of the electrode/cannula tip was reconstructed relative to the skull using stereo-photogrammetric techniques. This procedure enables high-resolution reconstruction (< 200um error) of electrode and cannula locations across experimental sessions (Cox et al., 2008; Issa et al., 2013a). Under assumptions of approximate planarity for the ventral surface of IT, we measured the distance between sites in IT using the Euclidean distance between X-ray reconstructed 3-D coordinates.

In total, we collected data for 25 inactivation experiments, with each inactivation experiment consisting of three consecutive behavioral sessions each, in two monkeys (monkey M: $n = 10$ experiments, monkey P: $n = 15$ experiments). Interleaved within this inactivation data collection, we additionally collected behavioral data for 18 control experiments, where each experiment again consisted of three consecutive control behavioral sessions each, with the same images and subtasks but with no injections, in both monkeys (monkey M: $n = 5$ experiments, monkey P: $n = 13$ experiments). These control data were used to estimate the natural variability in performance across behavioral sessions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral metrics

We previously introduced several metrics to characterize behavior in this pairwise object discrimination paradigm (Rajalingham et al., 2018). Here, we focus on the highest resolution behavioral metric that can be reliably measured in a single behavioral session, the one-versus-other object level performance metric (previously termed B.O2). Briefly, this metric is a pattern of pairwise object discrimination performances. Note that the distributions of images of each object were comparable across inactivation experiments, with the exact same 40 images/object across all sessions of experiment 1, and 100 images/object generated with the exact same generative parameters for experiment 2. For each pairwise object discrimination subtask, performance was estimated using a sensitivity index d' (Macmillan, 1993): $d' = Z(\text{hit rate}) - Z(\text{false alarm rate})$, where $Z(.)$ is the inverse of the cumulative Gaussian distribution. All d' estimates were constrained to a range of $[0, 5]$.

Recall that each inactivation experiment consisted of three behavioral sessions. We first equated the number of trials per session by selecting the first N trials of each session, where N was the minimum number of trials across the three sessions. For each of these three behavioral sessions, we then computed a pattern of performances across subtasks. We observed a small amount of variability in performance between the pre- and post-control behavioral sessions that was consistent with the effects of learning (see Figure S4A for quantification). To ensure that out measurements of the inactivation deficits were robust to this background variability, we defined the control behavioral performance as the average of the pre-control and post-control performances: $\psi_{control} = (\psi_{precontrol} + \psi_{postcontrol})/2$. To measure the behavioral deficit from inactivation, we estimated a behavioral deficit pattern (δ) as the difference between inactivated and control performance over subtasks:

$$\delta = \psi_{inactivated} - \psi_{control}$$

We additionally estimated a normalized behavioral deficit pattern as

$$\delta_n = \frac{\psi_{inactivated} - \psi_{control}}{\psi_{inactivated} + \psi_{control}}$$

For one set of analyses (Figure 6C), we characterized behavior using the one-versus-all object level performance metric (previously termed B.O1), corresponding to a pattern of discrimination performances per object. However, an important caveat is that the distributions of subtasks of each object were not comparable across all experiments, i.e., not all pairwise discrimination subtasks spanning these objects were tested.

We additionally characterized behavioral performance using several other behavioral metrics: balanced accuracy, choice bias, and reaction times (see Figure S2). However, we note that the choice of focusing on sensitivity (d') for this behavioral paradigm is principled and indeed was made prior to carrying out the experiments (Rajalingham et al., 2015, 2018). The overall behavioral performance (or accuracy) on a two-alternative forced choice task can be decomposed into sensitivity (measured by a sensitivity index d', see above) and bias (measured by a criterion index $c = 0.5(Z(\text{hit rate}) + Z(\text{false alarm rate}))$ (Macmillan, 1993). Importantly, measuring accuracy alone fails to disambiguate between these two separate components. In addition to this primary caveat, accuracy is bounded (i.e., sensitive to floor/ceiling effects) and non-linear with respect to the underlying representation in a signal detection framework (Green and Swets, 1966). In previous work, we observed that patterns of choice bias are much less reliable (across subjects) than sensitivity (Rajalingham et al., 2015). Here, we additionally observed that a given subject's choice bias can substantially vary under natural conditions (see Figure S2B). Finally, in this work, we did not explicitly train the animals on a speeded task (e.g., where faster response leads to more reward), and thus do not *a priori* expect any changes in reaction time. For these reasons, while we observe statistically significant inactivation effects with respect to accuracy and choice bias, we focus our primary claims on changes in sensitivity.

Sparsity of deficit

We quantified the non-uniformity of the behavioral deficits using a sparsity index $SI(x)$ (Vinje and Gallant, 2000) as follows:

$$A(x) = E[x]^2 \Big/ E\left[x^2\right],$$
$$SI(x) = (1 - A(x))/(1 - 1/N)$$

where $E[.]$ denotes the expectation of, and N is the length of the vector x. When applied to a behavioral deficit pattern with no sampling noise, $SI(\delta)$, this index has a value of 0 for perfectly uniform deficit patterns, and a value of 1 for perfectly one-hot deficit pattern. To ensure that the sparsity of the behavioral deficit did not purely reflect non-uniformity in the behavioral difficulty across subtasks, we additionally computed this index from the normalized deficit pattern vector $SI(\delta_n)$. We computed the SI for each inactivation site, and estimated the average across all sites.

To ground this empirical SI value in intuition, we estimated the corresponding SI distributions for different simulated behavioral deficit patterns with varying degrees of non-uniformity across subtasks, and with comparable sampling noise to that in our actual behavioral data (i.e., a finite number of trials). To estimate the expected SI distribution from a deficit with P% of subtasks affected, we performed the following simulation. For each inactivation site, we computed an estimate of the deficit pattern (δ) from a random bootstrap sample of trials. From this deficit pattern estimate, we set all but the top P% of deficit values to zero, and randomly shuffled the position of remaining non-zero entries. We averaged the resulting deficit pattern estimates across bootstrap samples to obtain a simulated deficit pattern with approximately equal, non-zero deficit on P% of subtasks. Finally, we computed the sparsity index for this simulated mean deficit pattern. By varying $P(= 10\%, 25\%, \ldots, 100\%)$, we obtained estimates of SI distributions expected from different degrees of non-uniformity across subtasks. Crucially, the simulated sparseness distributions preserved a number of key sources of variance — including the number of sites, the number of subtasks for each site, the performance of each subtask for each site, and the average performance deficit (across subtasks) for each site — because all these sources of variance were fixed for the empirical and simulated sparseness estimates, and the random shuffling were done after computing the behavioral deficits.

Neuronal readout models

To investigate the link between neuronal activity and behavioral deficits, we constructed and tested a number of decoding models (DMs). Each of these models predicts an inactivation pattern from the activity of neurons recorded in close anatomical proximity to the injection site. As described above, multi-unit neuronal activity was measured in response to the same images under a passive viewing paradigm and could thus be used as the input to each decoding model. We constructed a feature matrix R from the firing rate responses over images (averaged over repetitions) all local multi-unit sites — see below for definition of local. Each tested decoder model maps R to a behavioral deficit prediction Δ. The local neural discriminability and local population discriminability models we tested here were loosely inspired from population readout models of IT (Majaj et al., 2015). Note, however, that the current implementations do not include the remaining (non-local) IT population as inputs, as we did not have access to a larger sample of IT. The specific local decoder models we tested here were: *local neural response models* (DM1, DM2) predict largest deficits for subtasks with images that yielded largest response from the local neuronal sites, and *local neural discriminability models* (DM3, DM4, DM5) predict largest deficits for subtasks for which the local neural population was most discriminative, as measured by a linear classifier. Crucially, these linear classifiers were not directly fit to the target inactivation patterns, but rather trained to perform on the pairwise object discrimination subtasks. The resulting pattern of neural decode predicted performances over subtasks was converted to the pattern of predicted deficit by simply taking a negative (i.e., the highest performing subtask is the one that is predicted to be most reduced by inactivation of the neurons contributing to the decode). The details of these five models are as follows:

1. DM1 (mean neural response): The deficit for each subtask $\Delta_{i,j}$ is estimated as the (negative of) neural response to objects i,j, averaged over sites and images ($\langle R \rangle_{sites,images}$).
2. DM2 (weighted mean neural response): The deficit for each subtask $\Delta_{i,j}$ is estimated as the (negative of) neural response to objects i,j, averaged over sites and images after weighting each site by its overall discriminability w ($\langle wR \rangle_{sites,images}$).
3. DM3 (local neural response discriminability): The neural image responses averaged over sites, ($\langle R \rangle_{sites}$), is used as a single neural feature f to train and test a linear SVM. The deficit for each subtask $\Delta_{i,j}$ is estimated as the (negative of) the SVM performance (in units of d') to objects i,j, averaged over images.
4. DM4 (local neural discriminability, mass action): Each neural site's image response (R_{site_k}), is used as a single neural feature f to train and test a linear SVM. The deficit for each subtask $\Delta_{i,j}$ is estimated as the (negative of) the SVM performance (in units of d') to objects i,j, averaged over images, summed over all sites k.
5. DM5 (local population discriminability): The local neuronal image response (R), is used to train and test a linear SVM. The deficit for each subtask $\Delta_{i,j}$ is estimated as the (negative of) the SVM performance (in units of d') to objects i,j, averaged over images.

For each inactivation site, we defined its "local" neural population as all recorded multi-unit sites within a distance of θ_d from the inactivation site, where distances were estimated from the X-ray reconstructed electrode and cannula locations. In an effort to be robust to sparse neuronal sampling and X-ray reconstruction error, we fit this hyper-parameter θ_d to best predict the inactivation deficit patterns, rather than use a single fixed distance threshold across all inactivation sites. Specifically, we tested a range of θ_d values (from 1mm to 4mm, in steps of 0.25mm) and selected the value that yielded the most predictive model. We did not cross-validate this hyper-parameter optimization due to data limits (as the data have already twice been split prior to estimating decoding model consistency). Note, however, that θ_d was optimized separately for each decoding model, ensuring that no particular class of models was disproportionately benefitted. The resulting optimal distance thresholds corresponded to the approximate known spatial spread of muscimol in the cortical tissue ($\sim 2-3$mm). Furthermore, results are very similar in magnitude without this hyper-parameter optimization. To ensure that decoding model consistencies do not simply reflect spurious correlations, we additionally tested the consistency of decoding models constructed from distant neural populations. Briefly, for each tested inactivation site, we compared the true behavioral deficit pattern with that obtained by the the decoding model prediction from the most distant site.

Noise-adjusted correlations

We measured the similarity between two behavioral deficit patterns δ_1, δ_2 (e.g., between true deficit patterns and predictions from a model) using a noise-adjusted correlation (DiCarlo and Johnson, 1999; Johnson et al., 2002). For each behavioral deficit pattern, we split all independent raw observations (e.g., behavioral trials) into two equal halves and computed the behavioral deficit pattern from each half, resulting in two independent estimates of the deficit pattern. We took the Pearson correlation between these two estimates as a measure of the reliability of that behavioral deficit pattern, given the data, i.e., the split-half internal reliability. To estimate the noise-adjusted correlation between two deficit patterns, we compute the Pearson correlation over all the independent estimates of deficits from each, and we then divide that raw Pearson correlation by the geometric mean of the split-half internal reliability of each deficit:

$$\tilde{\rho}(\delta_1, \delta_2) = \frac{\rho_{\delta_1,\delta_2}}{\sqrt{\rho_{\delta_1,\delta_1} \times \rho_{\delta_2,\delta_2}}}$$

Since all correlations in the numerator and denominator were computed using the same amount of trial data (exactly half of the trial data), we did not need to make use of any prediction formulas (e.g., extrapolation to larger number of trials using Spearman-Brown prediction formula). This procedure was repeated 10 times with different random split-halves of trials. Our rationale for using a reliability-adjusted correlation measure was to account for variance in the behavioral deficit that is not replicable by the subtask condition. If two behavioral deficits are identical, then their expected noise-adjusted correlation is 1.0, regardless of the finite amount of data that are collected. The noise-adjusted correlation was used to compute the similarity between observed and predicted behavioral deficit patterns (e.g., for testing neural readout models), as well as for the similarity between two different behavioral deficit patterns arising from two different inactivation sites.

Statistical testing

Unless otherwise specified, we estimated the uncertainty in behavioral deficit measurements (i.e., delta, see above) via bootstrap resampling of trials, repeated 100 times. The standard error of each delta measurement was estimated as the standard deviation of its bootstrap distribution. For statistical tests, we performed one-tailed exact tests, by computing the empirical probability of observing a sample below zero. To compute this probability from the empirical bootstrap distribution, we fit a Gaussian kernel density function to the empirical distribution, optimizing the bandwidth parameter to minimize the mean squared error. This kernel density function was evaluated to compute a p value, by computing the cumulative probability of observing a positive behavioral delta.

DATA AND SOFTWARE AVAILABILITY

Code and data are available by request to the Lead Contact (dicarlo@mit.edu).