Neuron

Article

Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition

Highlights

- Reversible inactivation of vlPFC induced specific deficits in object recognition
- Induced IT population decode deficits were specific to "late-solved" images
- Deficits in object recognition behavior were higher for late-solved images
- vlPFC inactivation causes IT responses to better match feedforward models

Authors

Kohitij Kar, James J. DiCarlo

Correspondence

kohitij@mit.edu

In Brief

Kar and DiCarlo show that reversibly inactivating parts of macaque vlPFC results in selective object recognition deficits for specific images that most likely depend on recurrent computations. Their results implicate vlPFC, a recurrently connected circuit node, as critical to producing behaviorally sufficient object representations in the primate ventral visual stream.

Kar & DiCarlo, 2021, Neuron *109*, 1–13
January 6, 2021
https://doi.org/10.1016/j.neuron.2020.09.035

CellPress

Neuron

CellPress
OPEN ACCESS

Article

Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition

Kohitij Kar[1,2,3,*] 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 01239, USA
[2]Center for Brains, Minds and Machines, Massachusetts Institute of Technology, Cambridge, MA 01239, USA
[3]Lead Contact
*Correspondence: kohitij@mit.edu
https://doi.org/10.1016/j.neuron.2020.09.035

SUMMARY

Distributed neural population spiking patterns in macaque inferior temporal (IT) cortex that support core object recognition require additional time to develop for specific, "late-solved" images. This suggests the necessity of recurrent processing in these computations. Which brain circuits are responsible for computing and transmitting these putative recurrent signals to IT? To test whether the ventrolateral prefrontal cortex (vlPFC) is a critical recurrent node in this system, here, we pharmacologically inactivated parts of vlPFC and simultaneously measured IT activity while monkeys performed object discrimination tasks. vlPFC inactivation deteriorated the quality of late-phase (>150 ms from image onset) IT population code and produced commensurate behavioral deficits for late-solved images. Finally, silencing vlPFC caused the monkeys' IT activity and behavior to become more like those produced by feedforward-only ventral stream models. Together with prior work, these results implicate fast recurrent processing through vlPFC as critical to producing behaviorally sufficient object representations in IT.

INTRODUCTION

A goal of visual neuroscience is to identify and model the brain circuitry that seamlessly solves the challenging computational problem of rapid visual object categorization (DiCarlo and Cox, 2007; Riesenhuber and Poggio, 2000; Yamins and DiCarlo, 2016). Previous studies (Freiwald et al., 2009; Hung et al., 2005; Kar et al., 2019; Logothetis and Sheinberg, 1996; Majaj et al., 2015) show that the pattern of neural activity in primate inferior temporal (IT) cortex can explicitly represent visual object identities. However, current models of core object recognition fall short of fully explaining both primates' behavioral image-by-image difficulty patterns (Geirhos et al., 2017; Rajalingham et al., 2018) and the distributed population activity patterns of IT neurons (Kar et al., 2019).

These models primarily belong to the family of deep convolutional neural networks (DCNNs) with predominantly feedforward architectures. More recent models are beginning to implement recurrent architectures (Kubilius et al., 2019; Nayebi et al., 2018; Spoerer et al., 2017), but experimental data to guide their development are needed. Toward that goal, we have recently demonstrated (Kar et al., 2019) the critical role of putative recurrent signals available at the late phases of the image-evoked IT responses in enabling accurate core object recognition, at least for some images. That study also speculated that the lack of recurrent computations in the feedforward DCNN models might have led to its poor behavioral accuracy and poorer prediction of the late-phase IT responses. Nevertheless, which recurrent circuit motifs in the primate brain are most critical: within the ventral stream, within IT, top-down from regions downstream of IT (e.g., prefrontal cortex [PFC] and amygdala), or all of the above? Identifying these circuits and inferring their computational functions is critical in developing the next generation of models of the primate visual intelligence and behaviors such as core object recognition.

Kar et al. (2019) determined, for each tested image, the time when response patterns of the IT neuronal population could sufficiently account for the monkey's object recognition performance on that image, referred to as the object solution time (OST; one OST computed per image; Figure 1A). They also identified hundreds of images that critically relied on the early (90–120 ms) and late (150–180 ms) phases of the IT responses after image onset (Figure 1B). These results point to a targeted disruption strategy to test the aforementioned critical recurrent circuits. Specifically, if a particular recurrent circuit motif is critical for core object recognition, then its disengagement should (1) prevent the emergence of linearly decodable object identity information in the late phases of the IT responses, with little or no effect on the early phase, and (2) result in a reduction in behavioral performance for the "late-solved" images, with little or no effect

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

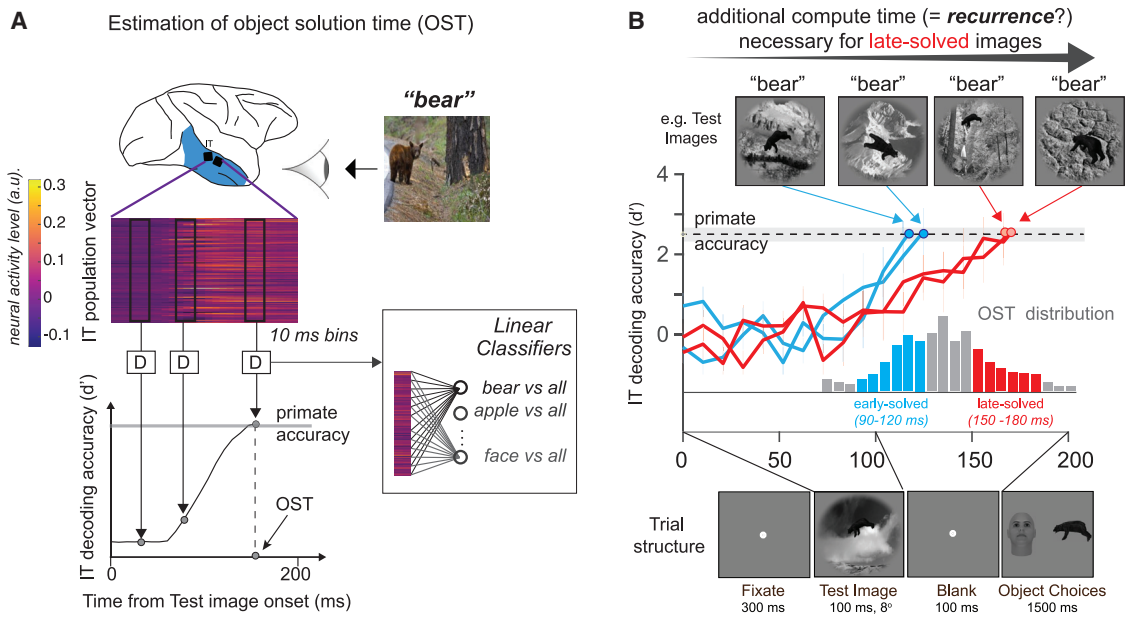

Figure 1. Motivation

(A) Estimation of object solution time (OST). For each image presentation (an example image of a bear is shown; 100 ms), we counted the number of multi-unit spike events (see STAR Methods for details) per site in nonoverlapping 10-ms windows after stimulus onset to construct a single population activity vector per time bin. These population vectors (image-evoked neural features) were then used to train and test cross-validated linear support vector machine decoders (D) separately per time bin. The decoder outputs per image (over time) were then used to perform a binary match to sample task (see STAR Methods) and obtain neural decode accuracies at each time bin. The time at which the neural decodes equal the primates' (pooled monkey) performance was then computed as the OST for that specific image.

(B) Temporal evolution of linearly decodable object identity information in IT on an image-by-image basis. For each tested image, we measured (Kar et al., 2019) the IT population response vector (n = 424 neural sites) across time (10 ms resolution). For each time point, we estimated the linear decodable information (cross-validated across images). Each image achieved a solution goodness (linear decode accuracy for object identity; d′) that matches the monkey's behavioral accuracy (average of d′ = 2.5 for the example images, shown as a gray shaded line) after different amounts of processing time (OST; gray histogram over; 1,320 tested images). Using a range of controls, Kar et al. (2019) concluded that images that exhibit longer OSTs (late solved; red curves show two examples) likely require more recurrent processing (relative to images that exhibit shorter OSTs, or early-solved images; blue curves show two examples).

on behavior performance for the "early-solved" images. In this study, we tested those two predictions for a circuit motif that is recurrently connected to the ventral visual stream, ventrolateral prefrontal cortex (vlPFC).

Among the multiple downstream targets of IT, we chose to first test vlPFC because (1) it is downstream of IT but has strong recurrent anatomical connections to IT (Borra et al., 2010; Webster et al., 1994; Yeterian et al., 2012); (2) following object-category learning, it has been shown to contain object-category selective neurons (Freedman et al., 2001, 2003) that maintain the category-related signals beyond the stimulus presentation, while these signals transiently disappear after the feedforward pass and then reappear later at the level of IT (suggesting a top-down feedback from vlPFC); (3) previous studies have demonstrated changes in IT resulting from lesion-based (Tomita et al., 1999), pharmacological (Monosov et al., 2011), and thermal perturbation (Fuster et al., 1985) of PFC; and (4) methods to silence vlPFC are experimentally straightforward, because vlPFC is downstream of IT. Specifically, we pharmacologically silenced (via muscimol, a $GABA^A$ agonist) ~0.4 cm^3 of vlPFC in each of two monkeys and measured changes in IT population activity at the multi-unit level (with chronically implanted Utah arrays ipsilateral to the targeted vlPFC; see Figure 3A) and the corresponding changes in core object recognition performance.

Our results show that the inactivation of vlPFC reduced the quality of the late-phase IT population activity, as assessed by the linear decodability of object identities. We also observed corresponding behavioral deficits in core object recognition tasks; the deficits were significantly higher for late-solved images. Interestingly, the inactivation of vlPFC caused the late-phase IT neural activity to become better explained by feedforward-only DCNN models of the ventral stream. These results argue that fast recurrent processing through vlPFC is critical to the production of fully robust object representation in IT and the core object recognition behavior that it supports and that current, feedforward-only computational models of the ventral stream lack these computations.

RESULTS

As outlined above, we reasoned that, if recurrent processing via vlPFC to the primate ventral stream is critical for robust core object recognition, then inactivating parts of the vlPFC should produce specific changes in the IT population activity patterns and specific behavioral deficits. In particular, the neural and

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

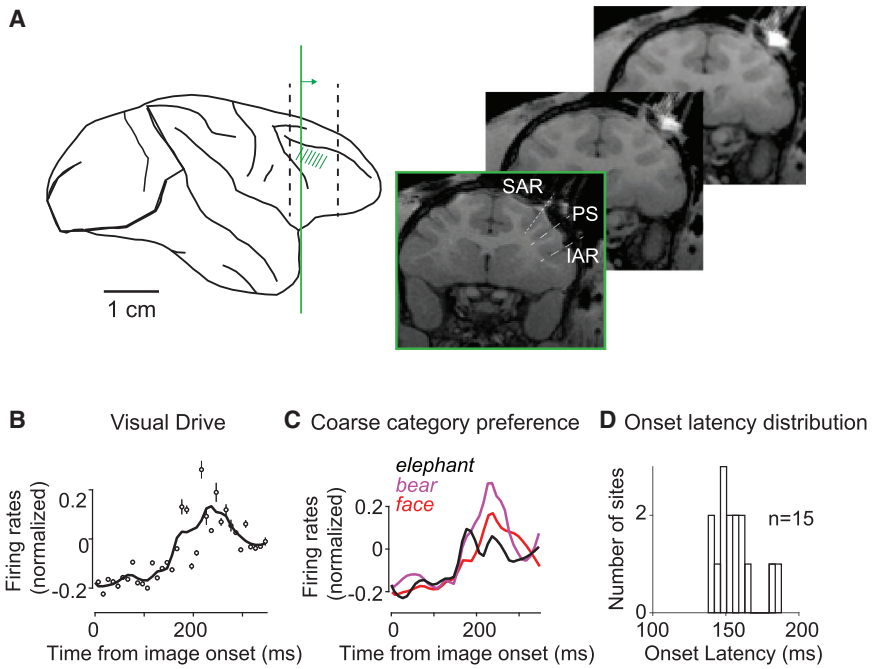

Figure 2. Approximate Location and Functional Properties of Injection Targets in vlPFC

(A) The left panel shows the approximate anterior-posterior (AP) boundaries (black dashed lines) of the chamber that was placed over vlPFC. The green line denotes the location of the coronal section displayed on the right. The arrow refers to more anterior locations matching the other coronal MRI images. The right panels show structural MRI images of the approximate targets of the muscimol injections (SAR, superior arcuate sulcus; PS, principal sulcus; IAR, inferior arcuate sulcus). Injections were made lateral and ventral to the principal sulcus (indicated by the green patch).

(B) Sample neural responses from a vlPFC site (averaged across 10 repetitions and 80 images) exhibiting visual drive. Error bars denote SEM across images.

(C) Coarse category selectivity of an example vlPFC neural site. Each curve is the average response per object category (8 images per category, 10 repetitions per image).

(D) Distribution of onset latencies of 15 neural sites in vlPFC (7 in monkey B, 8 in monkey N).

behavioral deficits should be higher for late-solved images, which do not produce a fully formed IT population representation until 150–180 ms after stimulus onset (Kar et al., 2019; see Introduction).

To test the role of vlPFC, we used pharmacological inactivation of subregions of vlPFC, as previously anatomically landmarked (Freedman et al., 2003; McKee et al., 2014; Tomita et al., 1999), and identified in this study by structural MRI (Figure 2A; see STAR Methods). Based on the expected locations of object-category-selective vlPFC neurons (Freedman et al., 2001, 2003), we first performed a single electrode measurement survey to locate vlPFC subregions that exhibited strong visual drive and coarse category selectivity (Figures 2B–2D; see STAR Methods). We then performed a second structural MRI (now with markers inserted at these locations) to ensure that the localized object-category selective vlPFC sites were anatomically consistent with previous reports (Freedman et al., 2001, 2003). We divided the data collection into two types of sessions: days with and without muscimol injections (Figure 3A). These two session types were repeated in an alternating sequence with at least 1 day of recovery after each muscimol session (Figure 3B; experimental timeline). This design may confound animal satiety and motivation with the effects of muscimol. However, the visual hemifield bias of our reported effects (see below) argues against that. In each session (day), monkeys performed the following tasks sequentially: a passive fixation task, a binary object discrimination task, and a second passive fixation task (see STAR Methods). On the second session (day), after the initial passive fixation task (included in the no-muscimol condition during all the analyses), we injected a total of 10 μL muscimol at five depths (2 μL each) separated by 0.5 mm in the previously localized vlPFC area (see STAR Methods for details). We injected in the left hemisphere of monkey B and right hemisphere of monkey N.

An Assay for Recurrence-Dependent Computation: Early-Solved versus Late-Solved Images

Previous studies (Hung et al., 2005; Majaj et al., 2015) have demonstrated that object identity is linearly expressed in the pattern of IT neural activity. Using linear decoders, we have previously estimated the precise time it takes for the macaque IT population to temporally evolve to this linearly explicit pattern for each of 1,320 images (Kar et al., 2019; briefly illustrated in Figures 1A and 1B). We refer to this time as the OST. OST is an estimate (done per image) of the amount of time needed to compute a behaviorally sufficient neural population solution in IT. Longer OSTs, therefore, suggest additional, putatively recurrent computations, beyond what could be achieved by the early, feedforward IT responses. In this study, our analyses primarily focus on comparing the neural and behavioral effects of vlPFC inactivation on the images that are solved quickly (early-solved images; OST range, 90–120 ms) with the effects on images that are solved slightly later (late-solved images; OST range, 150–180 ms).

As outlined in the Introduction, we hypothesized that if the inactivation of vlPFC (Figure 3A) disrupted behaviorally critical recurrent computations, then we should expect to see specific changes in IT population codes, and we should also see specific changes in behavior. In particular, we should observe a more significant muscimol-induced IT decode and behavioral performance deficit for images with late OSTs (H_2; Figure 3C, bottom panel). The other possibilities are that we observe no change (H_0; Figure 3C, top panel) in behavioral performance across images or an overall shift in the behavioral performance consistently across images with varied OSTs that

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

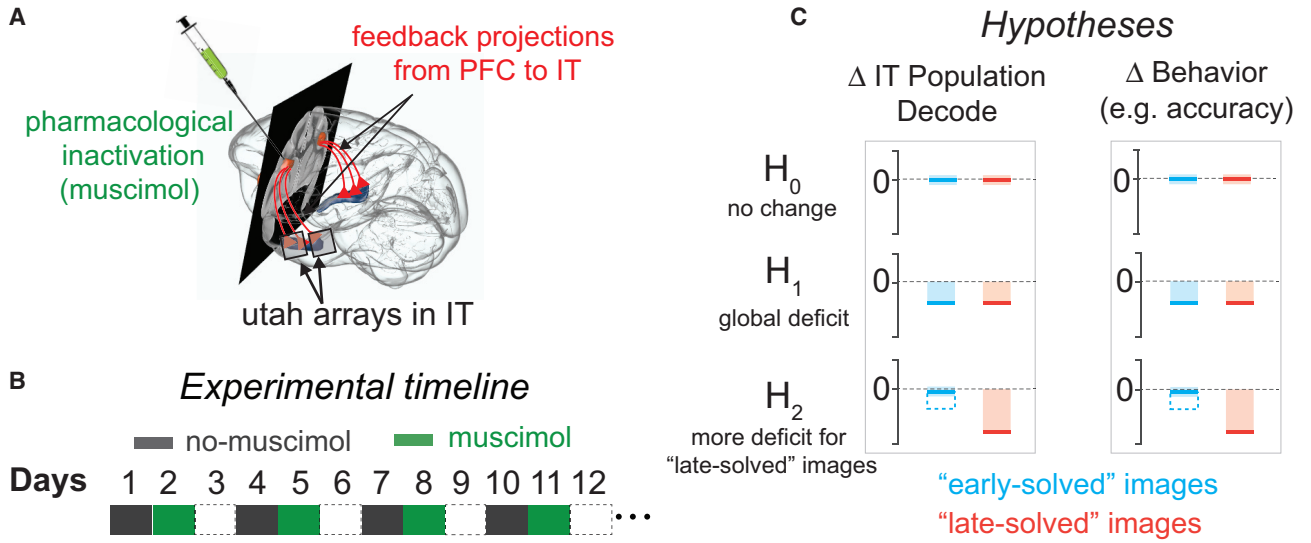

Figure 3. Experimental Setup and Hypotheses

(A) Pharmacological inactivation of vlPFC (ipsilateral to the IT recording location) with simultaneous IT population recordings.

(B) We divided the experiments into two different sessions, without (gray boxes) and with (green boxes) muscimol injections, conducted on consecutive days, with the exception of a passive viewing session before injections on the days with muscimol injections (see STAR Methods). We repeated each session in the same order after a minimum gap of 1 day (empty boxes). We completed at least 10 sessions for each condition type.

(C) Hypothesized effects of vlPFC inactivation. One hypothesis (H_0) is that the robustness of the IT object codes for core object recognition (~200 ms of processing) does not rely at all on vlPFC, which predicts no change in IT responses or behavior for both early- (blue bar) and late-solved (red bar) images. Another hypothesis (H_1) is that vlPFC plays an overall modulatory role in ventral stream computations, which predicts deficits in IT population solution goodness and behavior that are equal for both groups of images. Finally, a third hypothesis (H_2) is that vlPFC as a critical recurrence node in the brain circuitry for core objection, which predicts larger IT population solution deficits and larger behavioral deficits for late-solved images. A mixture of H_1 and H_2 is also possible (see open blue bars; see Discussion for alternative interpretations).

might indicate a global shift in arousal (H_1; Figure 3C, middle panel).

vlPFC Inactivation Reduces IT Late-Phase Population Activity

We first explored the effect of vlPFC inactivation on the quality of the IT neural population patterns evoked by each image. Upon visual inspection (Figure 4B), we observed that vlPFC inactivation did not produce a reduction in the mean initial (90–120ms) image-driven activity. However, vlPFC inactivation appeared to moderately reduce the later portion of the IT responses (i.e., starting ~140 ms after image onset). To look more closely, we compared IT responses at two specific time bins, early phase (90–120 ms; Figure 4C) and late phase (150–180 ms; Figure 4D). We found that across the entire recorded IT population (n = 153 sites), vlPFC inactivation produced no significant difference in the mean response (averaged over all images) in the early phase (ΔR^{early} = −18% ± 46.4%, mean ± SEM; paired t test; t(152) = 0.5885, p = 0.5571). However, vlPFC inactivation produced a significant reduction in mean late-phase IT responses (averaged across images; ΔR^{late} = −31.83% ± 10.4%, mean ± SEM; paired t test; t(152) = 8.5906, p < 0.0001). Also, we noted that the time of the emergence of a drop in the mean IT response (black versus green line in Figure 4B) coincided with the latencies of the vlPFC neurons that we recorded at the targeted injection sites (refer Figure 2D) as well as previously measured latencies of neurons in this area (Freedman et al., 2001, 2003). We note that these mean firing rate effects are also consistent with prior causal perturbation studies in other pairs of visually driven cortical areas (see Discussion).

vlPFC Inactivation Selectively Disrupts the Late-Phase IT Population Code

The neural representations that enable robust object recognition are more subtle than the mean firing rates analyzed above. Indeed, we previously reported that while many images evoke high mean firing rates in IT cortex, a linearly readable solution of the foreground object in those images is not present in that activity and emerges only later after subtle changes in the neuron-by-neuron distributed population code (Kar et al., 2019). Thus, we next aimed to examine the temporal evolution of the quality of the IT population code for early-solved versus late-solved images. Here, we assessed "quality" as the ability of the population code to support a linear readout of object identity for held-out test images (i.e., via cross-validation; see STAR Methods). As outlined before, we sought to specifically test the hypothesis (H_2; Figure 3C, left column) that vlPFC feedback to the ventral stream, is particularly critical to the development of late-phase IT object solutions. This hypothesis predicts that vlPFC inactivation should induce more significant disruptions in the quality of the IT population code for late-solved images compared to the early-solved images at their corresponding OSTs. To control for the behavioral accuracy levels across images, we sub-selected images (out of the total 1,320 tested images) for two

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

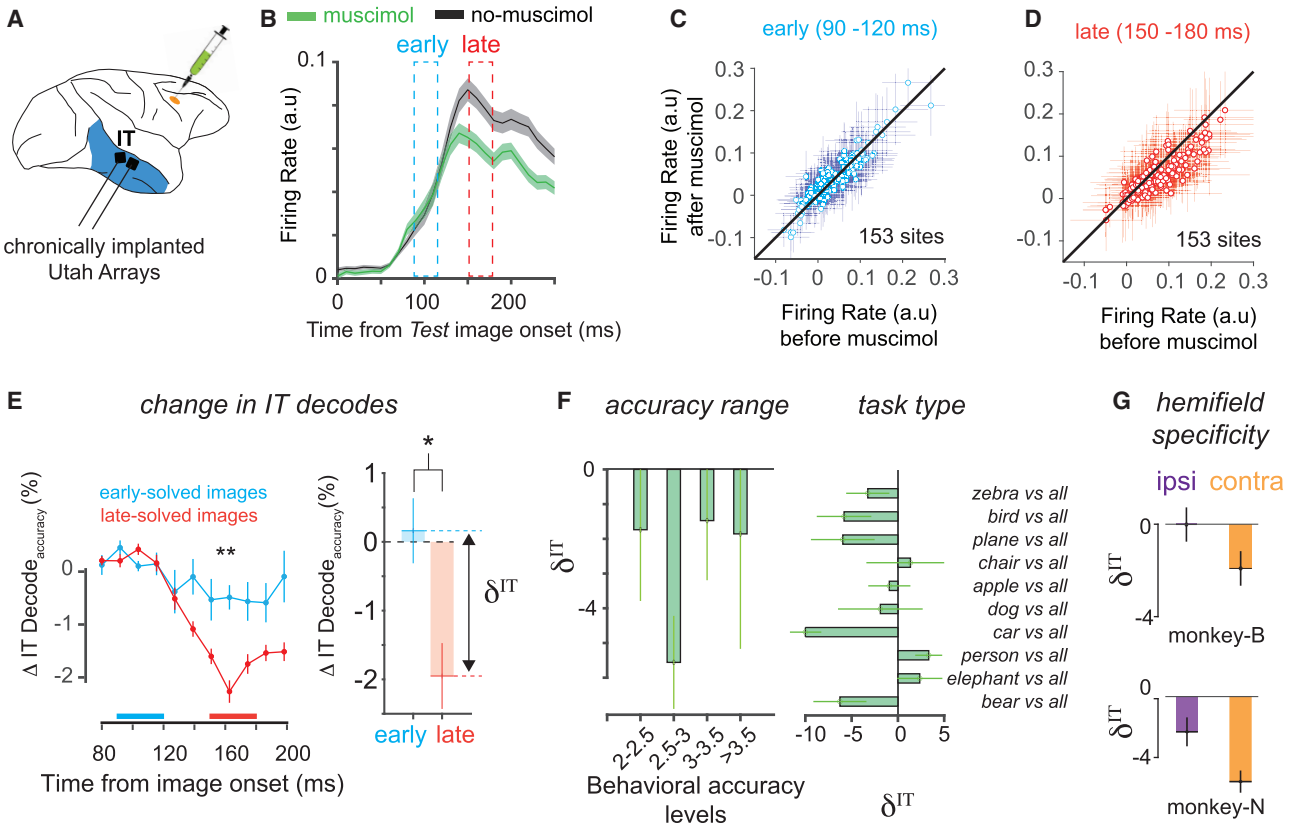

Figure 4. Neural Experiments and Results

(A) We measured neural responses from 153 sites in the IT cortex across two monkeys while they performed a battery of core recognition tasks, with and without muscimol injections in vlPFC.

(B) Normalized mean IT firing rate in the two conditions (black, no-muscimol control condition; green, after muscimol injections in vlPFC). The shades indicate SEM across images.

(C) We observed no significant differences across neurons at the early phase (90–120 ms) of the IT responses (ΔR^{early} = −18% ± 46.4%, mean ± SEM; paired t test; t(152) = 0.5885, p = 0.5571).

(D) We observed a small but significant reduction in firing rates at the late phase (150–180 ms) of the IT responses (ΔR^{late} = −31.83% ± 10.4%, mean ± SEM; paired t test; t(152) = 8.5906, p < 0.0001). Error bars for (C) and (D) denote the standard deviation of responses across images per neuron.

(E) Images (n = 234) with late OST (in red) showed a significantly higher drop in IT population decode accuracy across time (left panel) and at their corresponding OST (right panel) upon vlPFC inactivation compared to the images (n = 209) with early OST (in blue). This comparison was made with all images that had a measured (behavioral) d′ between 2 and 4, as measured in separate animals (Kar et al., 2019). Error bars denote SEM across images. We quantified the strength of this interaction as the difference in the muscimol-induced change (right panel), and we refer to that measure as δ^{IT}.

(F) The mean δ^{IT} was consistently less than 0 for images selected in different ranges of behavioral accuracies. We also observed a negative trend for most, but not all, recognition sub-tasks (t test, t(9) = 1.9718, p = 0.0401). Error bars denote bootstrap confidence interval (CI) (95%).

(G) Interaction strength was significantly stronger when we restricted the measurements to images where the object center was in the contralateral visual field (monkey N, ipsilateral δ^{IT}: −2%, contralateral δ^{IT}: −5.8%, permutation test of difference, p < 0.001; monkey B, ipsilateral δ^{IT}: 0.1%, contralateral δ^{IT}: −2%, permutation test of difference, p < 0.001). Error bars denote bootstrap CI (95%).

See also Figure S1.

groups, early solved (209 images) and late solved (234 images), that all had a (pre-muscimol) d′ between 2 and 4 (as measured in an earlier study; Kar et al., 2019).

First, we observed that the quality of IT neural population codes (as estimated by linear decode accuracies of object identity) were significantly less accurate at later time points after vlPFC inactivation (>150 ms after image onset; median reduction = −2.44%, t test, t(441) = 5.11, p < 0.001; Figure S1A). Furthermore, to estimate whether the muscimol-induced change in IT linear decodability of objects was dependent on the previously estimated OST values (Kar et al., 2019), we compared the IT decode accuracies for the early-solved and late-solved images at each time point (10-ms nonoverlapping bins) after the onset of the test image stimulus (Figure 4E, left panel). We observed that although there is a small but significant drop in the late-phase IT decodes for early-solved images (blue curve), this drop is significantly less than that for the late-solved images (comparison of blue and red curves between 150 and 180 ms; t test, t(441) = 7.3; p < 0.0001). To directly test the effects of vlPFC inactivation at the most behaviorally relevant IT response times, we focused primarily on the corresponding OSTs per image for future analysis (Figure 4E, right panel). We refer to the

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

difference (early minus late) in the muscimol-induced deficits as δ^{IT} (as shown in Figure 4E). We observed that vlPFC inactivation disrupts the formation of IT solutions for the late-solved images more than it disrupts the formation of IT solutions for the early-solved images (ΔIT population $decode_{accuracy}^{early}$ = 0.16% ± 0.53% ; ΔIT population $decode_{accuracy}^{late}$ = −2% ± 0.61%, median ± SEM ; t test, t(441) = 2.4084, p = 0.0165; Figure 2E). Moreover, we found that this effect persisted even with different behavioral accuracy level choices (behavioral levels considered in d′: < 2, 2–2.5, 2.5–3, > 3; corresponding δ^{IT} values were −1.63%, −5.63%, −1.57%, and −1.9%; Figure 4F). Also, δ^{IT} was significantly less than zero considering each of the 10 tested objects (10 tasks, t test, t(9) = 1.9718, p = 0.0401; Figure 4F). We observed that the δ^{IT} values, when measured separately for each monkey, were significantly more negative for images where the object center was present in the contralateral hemifield (monkey N, ipsilateral δ^{IT}: −2.01%, contralateral δ^{IT}: −5.8%, permutation test of difference, p < 0.001; monkey B, ipsilateral δ^{IT}: 0.1%, contralateral δ^{IT}: −2%, permutation test of difference, p < 0.001; yellow bars; Figure 4G) compared to those in the ipsilateral hemifield (purple bars; Figure 4G). We also observed that if we grouped images based on the latency of IT decodes estimated in the current study, images that take longer to reach a specific threshold (accuracy of 0.6) showed higher decoding deficits upon vlPFC inactivation (see STAR Methods for details). Taken together, our results demonstrate that vlPFC inactivation disrupts the formation of IT neural population solutions more strongly for images for which those solutions take longer to develop, consistent with the hypothesis that vlPFC is part of the critical recurrent circuitry.

vlPFC Inactivation Produces Larger Behavioral Deficits for Late-Solved Images

To further test how vlPFC inactivation affects behavior across late-solved and early-solved images, we measured the monkeys' behavioral performance during an array of binary object discrimination tasks (Figure 5B). Identical to Kar et al. (2019), in each image, the primary visible object belonged to one of 10 different object categories (Figure 5A). First, we observed that there was a significant overall reduction (Δperformance = 6.03% ± 0.3% [mean ± SEM], paired t test; t(859) = 17.13, p < 0.0001; Figure S3A) in performance for a binary object discrimination task (Figure 5B) across all sessions after the muscimol injections. Consistent with hypotheses H_2 (Figure 3C), vlPFC inactivation caused a significantly higher reduction in performance for late-solved images compared with early-solved images ($\Delta performance^{early}$ = −4.76% ± 0.45%; $\Delta performance^{late}$ = −7.4% ± 0.5% [median ± SEM]; t test, t(441) = 2.3978, p = 0.0085) . We refer to the difference in the behavioral deficits for the early versus the late-solved images as δ^{B} (as shown in Figure 5C). We observed that δ^{B} was consistently negative (i.e., greater behavioral deficits for late-solved images compared to early-solved images) across images grouped according to different behavioral accuracy level choices (behavioral levels considered in d′: < 2, 2–2.5, 2.5–3, > 3; corresponding δ^{B} values were −4.24%, −1.9%, −2.23%, and −1.9%; Figure 5D). Also, δ^{B} was significantly less than zero considering each of the ten tested objects (10 tasks, t test, t(9) = 2.6245, p = 0.0276). We observed that the δ^{B} values when measured separately for each monkey were significantly higher for images in which the object center was in the contralateral hemifield (monkey N, ipsilateral δ^{B}: −2%, contralateral δ^{B}: −4.8%, permutation test of difference, p < 0.001; monkey B, ipsilateral δ^{B}: 0.1%, contralateral δ^{B}: −3.3%, permutation test of difference, p < 0.001; Figure 5E, yellow bars), compared to those in the ipsilateral hemifield (Figure 5E, purple bars). We also observed an overall increase in reaction times after muscimol injections (ΔRT = 46.3 ± 2.1 ms; t test; t(858) = 16.3729, p < 0.001). Similar to the behavioral accuracy results, we observed that vlPFC inactivation increased reaction times, and that this increase was significantly higher for late-solved images than for early-solved images (ΔRT^{early} = −34 ± 4.19 ms ; ΔRT^{late} = 55 ± 3.9 ms, median ± SEM ; t test, $t(441) = 2.0488$, p = 0.04; Figure S3C).

These results show that core object discrimination behavior in macaques is disrupted by the inactivation of vlPFC, establishing this area as a critical component of the brain circuitry that is involved in core object recognition. Furthermore, the performance changes (deficits in task accuracy and reaction time) depended on the image being processed; the deficits were more severe for images that more likely depend on recurrent processing (as indexed by each image's IT OST; Kar et al., 2019). These behavioral results are qualitatively consistent with the IT neural results (Figure 4) under the assumption that the behavior is the consequence of mechanisms that are approximated by linear readouts from IT (Majaj et al., 2015). The latter assumption is further strengthened by our observation that the capability of the current best decoding model (Majaj et al., 2015; see STAR Methods) linking IT population activity to trial by trial monkey behavior (see Figure S5) did not change significantly after the inactivation of vlPFC.

Inactivation of vlPFC Causes the Ventral Stream to Operate More Similarly to Feedforward Computational Models

We have previously shown (Kar et al., 2019) that some feedforward DCNNs (specific DCNNs) predict the early feedforward responses of the IT neurons quite well but are far worse at predicting the late-phase IT responses. These prior results (and other work; see Discussion) suggest that the early-phase IT responses are primarily the product of feedforward computations but that the late-phase IT responses are a more balanced mixture of feedforward and recurrent computation (e.g., through vlPFC, as suggested by the results above). Under this hypothesis, the relatively weak ability of these DCNN ventral stream models to explain the late-phase IT responses is due to the lack of the appropriate recurrent computations. If we assume that vlPFC inactivation removes those additional recurrent computations (or blocks the transmission of the results of those computations to IT), vlPFC inactivation should make the late-phase IT representations revert to a more feedforward-only mode of operation. vlPFC inactivation should, therefore, make the top of the ventral stream operate more like a feedforward-only network.

To test this, we used a set of existing feedforward DCNN models (refer Table S1), and we asked whether vlPFC inactivation causes the late-phase IT response (recorded during a passive fixation task) to become better explained (predicted) by these

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

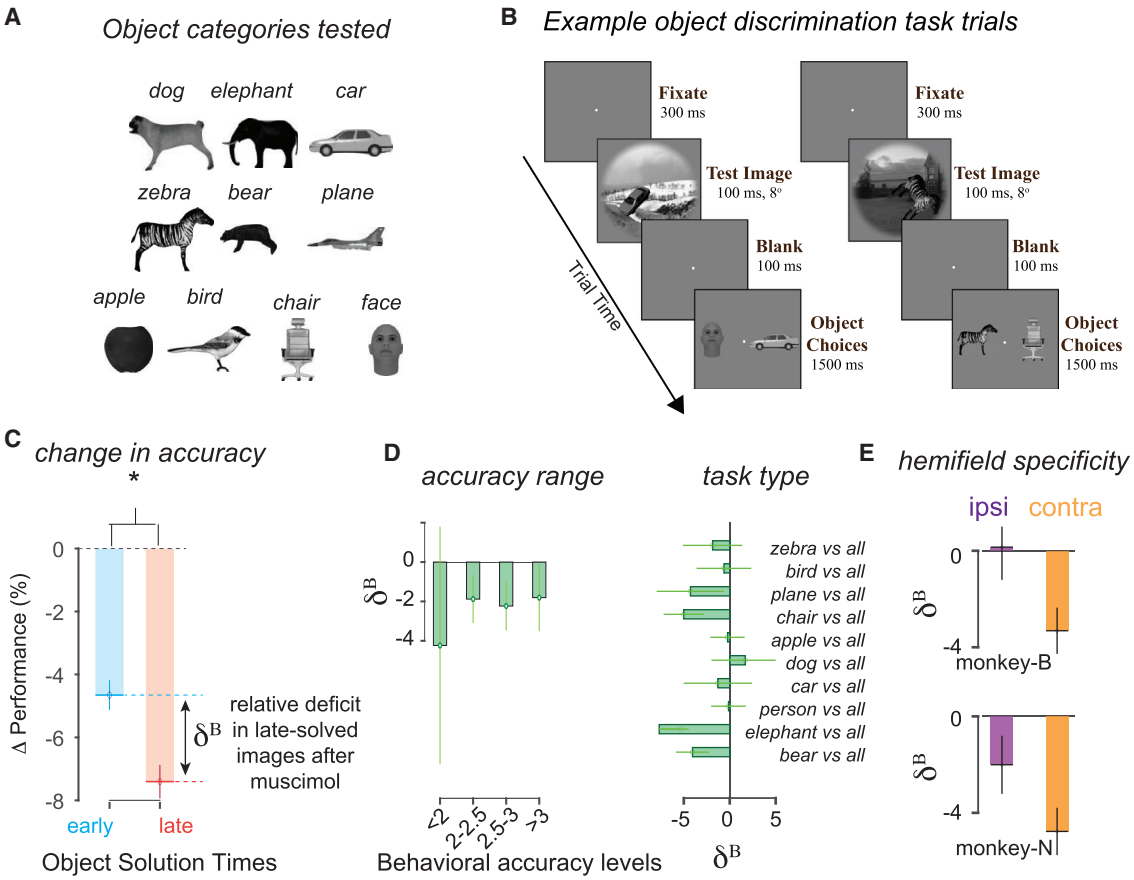

Figure 5. Behavioral Experiments and Results

(A) We tested behavioral performance on ten object categories, where performance was derived from the corresponding 45 binary object discrimination tasks with those 10 categories.

(B) Two example trials of the binary object discrimination task showing the timeline of events. Monkeys fixate on a central dot, and then the test image at 8° containing 1 of 10 possible objects is shown for 100 ms (shown is a car [left trial] and a zebra [right trial]). After a 100-ms delay, a canonical view of the target object and a distractor object (one of the other nine objects) appears (randomly assigned on each trial to the left and right positions), and the monkey indicates which object was present in the test image by making a saccade to one of the two choices. We compared performance on sessions with and without muscimol injections in vlPFC.

(C) vlPFC inactivation resulted in a larger performance drop among images (n = 234) with late OST (red bar), compared to the images (n = 209) with early OST (see Results for statistics). This comparison was made with all images that had a measured d′ between 2 and 4. Error bars denote SEM across images. We quantified the strength of this interaction as the difference in the muscimol-induced change, and we refer to that measure as δ^B.

(D) We observed that the mean δ^B was consistently less than 0 for images selected in different ranges of behavioral accuracies. We also observed a negative trend for most, but not all recognition sub-tasks (t test, t(9) = 2.6245, p = 0.0276). Error bars denotes bootstrap CI (95%).

(E) We found that the interaction strength was significantly stronger when we restricted the measurements to images where the object center was in the contralateral visual field (monkey N, ipsilateral δ^B: −2%, contralateral δ^B: −4.8%, permutation test of difference, p < 0.001; monkey B, ipsilateral δ^B: 0.1%, contralateral δ^B: −3.3%, permutation test of difference, p < 0.001). Error bars denotes bootstrap CI (95%).

See also Figure S3.

feedforward models (Figure 6A). We used standard measures of mapping the components of feedforward models onto the responses of individual IT neural sites (Kar et al., 2019; Schrimpf et al., 2018; Yamins et al., 2014; see STAR Methods), and we took the goodness of fit to be the median predictivity across all recorded neural sites. Remarkably, we observed that vlPFC inactivation significantly improved the match of the late-phase (150–180 ms) IT responses to the feedforward DCNN (AlexNet "fc7") predictions (median late-phase %EV: without muscimol = 21.98%, with muscimol = 28.28%; paired t test across neurons; t(152) = 8.55, p < 0.0001; Figure 6B, top panel; also see Figure S4B). Consistent with this, we also found that the vlPFC inactivation caused the late-phase IT responses to be more similar to the early-phase IT responses, as measured by correlation of image response rank order (early versus late) compared across the muscimol and no-muscimol conditions (paired t test; t(152) = 7.24; p < 0.001, see STAR Methods). These results were also consistent across multiple feedforward models (see Figure S5C) and across the neural data measured during the object discrimination task (median ΔIT predictivity was 5.21% ± 1.8%).

We also know from previous work (Rajalingham et al., 2018) that the DCNN models of core object recognition fail to explain

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

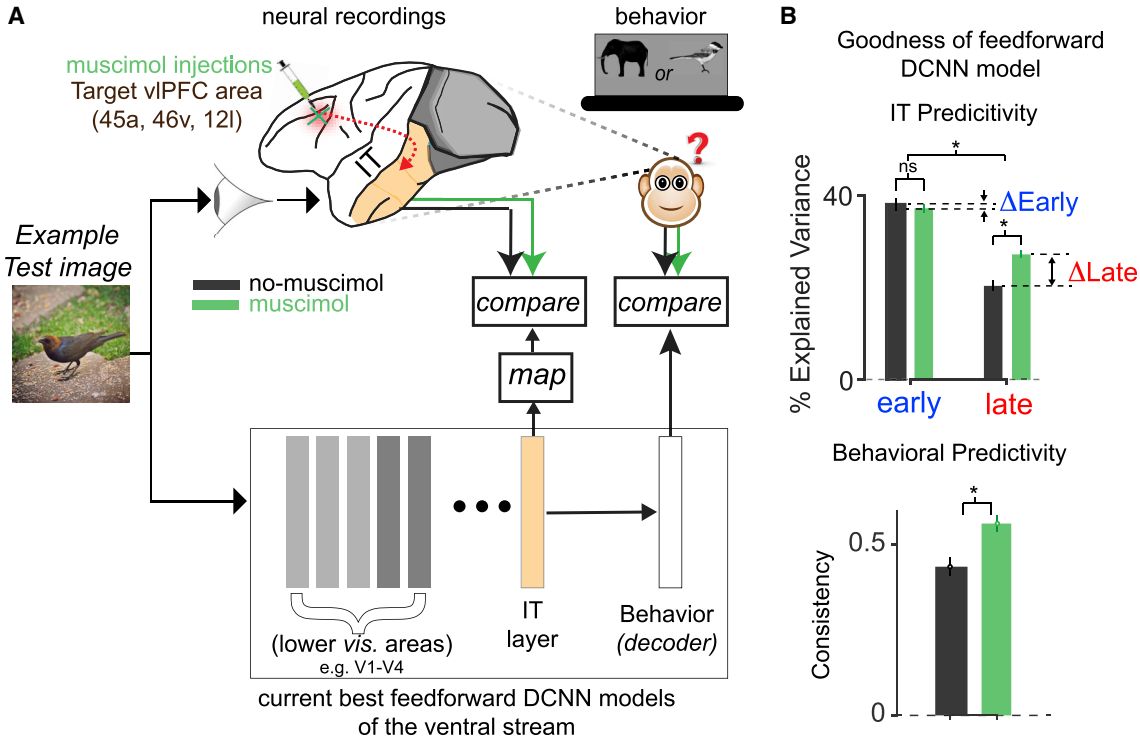

Figure 6. Comparison with Computational Models: vlPFC Inactivation Causes the Ventral Stream to Behave More Like Feedforward Models

(A) We showed 683 images to the monkey (fixated passive viewing) while recording simultaneously from their IT cortex, with and without vlPFC inactivation (top panel). The dashed red line denotes the recurrent pathway between vlPFC and the primate ventral stream. We compared the IT responses with and without vlPFC inactivation to those of the penultimate ("IT") layers of a feedforward DCNN model of the ventral stream (bottom panel) using previously established methods (see STAR Methods). We also compared the pattern of monkeys' behavioral responses (pattern of difficulty over images, see Rajalingham et al., 2018) with and without vlPFC inactivation to the model's behavioral pattern. In both types of comparisons, the key measure is referred to as predictivity, as it assesses the goodness of model predictions on new images.

(B) Top: comparison of IT predictivity (%EV) of AlexNet (fc7) for early (90–120 ms) and late (150–180 ms) responses, without (black) and with (green) vlPFC inactivation. We observed that vlPFC inactivation resulted in a significant increase in the match of the late phase of the IT population pattern to the feedforward DCNN "IT" population pattern. No significant changes were observed for early responses. Error bars denote SEM across 153 neural sites. Bottom: vlPFC inactivation resulted in a slight but significant increase in the match of the monkeys' object recognition behavior to the object recognition behavior of the feedforward models (match assessed as the correlation (model versus monkey) of the image-by-image pattern of difficulty).

See also Figures S4 and S5 and Table S1.

primate behavior at an image-by-image level fully; that is, those models do not fully explain and predict which images primates perform well on and which images they perform poorly on. Our results show that vlPFC inactivation changed the late-phase IT population patterns such that they became better matched to the "IT" (penultimate) layers of feedforward DCNNs. Taken together with prior work that tightly linked primate object recognition behavior to patterns of IT population activity (Majaj et al., 2015), we asked whether the vlPFC inactivation also changed the monkey image-by-image behavior patterns. Indeed, we observed that vlPFC inactivation significantly, improved the image-by-image consistency (normalized correlation, see STAR Methods) between the monkeys' object recognition behavior and the object recognition behavior of the feedforward models. (behavioral predictivity without vlPFC inactivation = 0.43, behavioral predictivity after vlPFC inactivation = 0.56; permutation test of difference; $p < 0.0001$; Figure 6B, bottom panel).

We have previously developed a state-of-the-art neural network model of the ventral stream named CORnet-S that contains some recurrence (Kubilius et al., 2019). That model includes areas V1, V2, V4, and IT, but it does not include vlPFC, and all of its modeled recurrences are local to each cortical area. We compared the similarity of its IT and behavioral layers with the primate IT and behavioral data (similar to the metrics used for the feedforward DCNNs) to ask how this model is related to the current experimental results. We found that as with the feedforward-only models, vlPFC inactivation caused the primate IT and behavior to increase (% Δ EV = 2.41% ± 0.47%, Δ behavioral predictivity = 0.14) its match to CORnet-S, suggesting that CORnet-S also does not contain all the normally functioning vlPFC recurrent processing. However, we also found that this increase was lower than that observed for shallower feedforward models and was similar in magnitude to the much deeper DCNNs (Figure S5B). This similarity in brain matching between very deep DCNNs and shallower recurrent networks like CORnet-S was previously reported (Kar et al., 2019) and followed from earlier theoretical work (Liao and Poggio, 2016). These model comparison results suggest that CORnet-S is

more like the ventral visual stream than feedforward-only models (as previously described in Kubilius et al., 2019), but the inactivation results presented here demonstrate the need to further improve the CORnet family of recurrent models by incorporating a recurrently connected vlPFC node.

In sum, these results suggest that vlPFC is a critical circuit node that is recurrently modulating the population dynamics of IT. Partially inactivating that node restricts the IT population pattern from correctly evolving away from its initial feedforward response pattern, leaving both the early (especially) and the late IT population patterns reasonably well approximated by current feedforward DCNN models of the ventral stream.

DISCUSSION

In this work, we investigated whether the recurrent circuit connecting macaque vlPFC to the ventral visual pathway is critical for executing robust core object recognition. We reasoned that if this bidirectional circuitry is indeed critical, then silencing parts of it should produce deficits in the quality of population activity recorded in the IT cortex that is responsible for accurate core recognition behavioral performance. More specifically, based on our prior work (Kar et al., 2019), we hypothesized that we should observe larger deficits for images that take slightly longer to solve and thus their solutions are more likely dependent on recurrent computations (late-solved images; benchmarked earlier in Kar et al., 2019).

Consistent with this hypothesis, we observed that vlPFC inactivation produced deteriorations in the quality of the IT population code and deteriorations in behavioral performance that were significantly higher for the late-solved images than for the early-solved images. Furthermore, we found that vlPFC inactivation caused the late phase of the IT population response and the monkey behavior to more closely match the "IT" and behavioral responses of some of the leading feedforward models of the ventral stream. These results suggest that vlPFC is part of a recurrent circuit that boosts the performance of the ventral stream (relative to shallow feedforward DCNNs) by reshaping the initial (early-phase, putatively feedforward-only) neural representations in the IT cortex, resulting in corresponding behavioral gains. Consistent with this, removal of vlPFC made the ventral stream operate more like a shallow feedforward system. When considered alongside prior work (Kar et al., 2019), this vlPFC circuitry is most critical for images that are challenging for shallow feedforward computer vision systems.

Experimental Guidance on Developing New Scientific Hypotheses of Ventral Stream Function

Our current best understanding of neural processing along the ventral stream is carried by specific models in the class of feedforward deep artificial neural networks. These models are the current best scientific hypotheses of the ventral stream, because they have the highest overall prediction accuracy (a primary test of a scientific hypothesis: Hempel, 1966; Popper, 1959) for image-evoked responses at all levels of the ventral stream (mean accuracy in V1, V2, V4, and IT; Schrimpf et al., 2018). However, because these models do not perfectly predict the image-evoked neural responses of these different areas of the ventral stream (for comparison across different models, see Schrimpf et al., 2018), multiple groups are working to develop even more accurate scientific hypotheses (e.g., Kubilius et al., 2019; Nayebi et al., 2018; Spoerer et al., 2017). What components do these current models lack? Clearly, the models are missing many things at the single-"neuron" level, such as voltage-gated channels to generate spikes, dendritic trees, and synaptic components. We motivated this study by first asking what critical network-level components are missing from these models.

Many studies and reviews have suggested the importance of including recurrent circuits to improve such models (Kar et al., 2019; Kietzmann et al., 2019; Lehky and Tanaka, 2016; Tang et al., 2018). This idea is motivated on both anatomical and functional grounds. For example, previous reports (Sugase et al., 1999) have demonstrated that different forms of information can be decoded from early and late responses in IT, suggesting a potential role of intra-areal recurrent inputs to shaping IT population response dynamics. Consistent with the hypothesis that recurrent signals modify late-phase IT population responses, Kar et al. (2019) showed that the ability of feedforward DCNNs to predict the IT population pattern significantly worsened as the IT response pattern evolved. They also showed that this latter portion of the IT population response pattern carries the linearly available object identity information for many specific images that enable primates to successfully solve them, vastly outperforming shallow feedforward DCNN computer vision models. In sum, the late phase of the IT population response is likely important for robust core recognition behavior, likely depends on recurrent circuits, and is largely missing from the current best models of the ventral stream. Thus, to produce models of the ventral stream that more closely mimic the mechanisms of the primate brain, a proper form of recurrent network-level processing is needed.

What type of recurrent processing is needed? To begin to answer that question, we started with an even more basic question: what circuit nodes in the brain are computing and carrying the recurrent signals that we see manifesting as a temporal evolution of the IT late-phase responses? Prior work suggests many potential sources of such signals, including within ventral stream bidirectional pathways, as well as top-down feedback from multiple downstream areas, including vlPFC, peri-rhinal cortex, amygdala, and striatum (for review, see Kravitz et al., 2013). For reasons outlined in the Introduction, in this study, we have specifically focused on vlPFC.

To test the functional importance of a downstream node that is recurrently connected to a target region of interest, many previous studies in the visual system (Bullier et al., 2001; Hupé et al., 1998; Sandell and Schiller, 1982; Wang et al., 2000) have used an inactivation method similar the one deployed here. In general, those studies report that this downstream manipulation results in a decrease in responses of neurons in earlier cortical areas, which is analogous to the reduction (~30%) in IT activity level that we have observed here (Figure 4B). For instance, inactivation of area MT (feedback node) via cooling led to a ~20%–40% decrease in V1 and V2 responses (refer to Figures 1 and 2 in Hupé et al., 1998). Focusing specifically on vlPFC and IT, prior studies have confirmed that IT responses, similar to other visual areas are modulated by feedback from downstream

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

areas. For example, Fuster et al. (1985) showed that temporary lesions produced by cooling in dorsolateral PFC affected color selectivity in IT neurons. Tomita et al. (1999) performed anterior and posterior commissurectomies and observed that the responses of IT cortical neurons are modulated by input from the prefrontal cortices, especially for visual information in the contralateral visual field.

Our work is consistent with those studies in that IT responses can be altered by vlPFC. However, unlike the work presented here, those earlier studies did not specifically investigate the changes in the distributed IT population code or primate behavior with respect to object recognition, which can guide the development of new models of primate vision. Specifically, that prior work did not engage on questions of the quality of information for recognition behavior at an image-by-image resolution or the differential importance of recurrent signals from vlPFC as measured in the early versus late responses of the IT population. Because of this, prior work could not distinguish between an overall modulatory role (H_1) and a specific set of recurrent computations (similar to H_2). To our knowledge, the current study is the first to causally test the necessity of the vlPFC to ventral stream recurrent circuit at such fast (<200 ms) but natural timescales, with simultaneous large-scale neural and behavioral measurements. Here, we have leveraged our previous findings (as reported in Kar et al., 2019) to employ a targeted disruption strategy for identifying critical recurrent circuits using predefined challenge images (that take additional solution times in IT). Therefore, our results provide evidence that feedback from vlPFC does not simply modulate IT (e.g., gain); it specifically improves the format of the distributed IT population code, and those improvements are specific to the late phase of this code.

However, the results reported here do not identify the exact circuitry involved in the reentry of information from vlPFC into the ventral stream. Previous anatomical studies have shown that the feedforward projections that connect the ventral stream to the prefrontal cortices originate in the anterior portions of the lower ventral bank and fundus of the STS (for a review, see Kravitz et al., 2013) and mainly target areas 45A/B, 46v, and 12r/l in vlPFC. On the other hand, feedback projections from these same areas in the PFC are distributed across the IT cortical areas TEO and TE (Gerbella et al., 2010). There is not much evidence of direct connections between these areas in the PFC and earlier visual areas (V1, V2, and V4), but we cannot rule out the possibility of indirect connections to the lower visual areas via the frontal eye fields and other regions.

Each of these possible circuit motifs is a hypothesis that must be, in the future, implemented as a set of neural network models for future experimental testing. Our neural measurements (with and without vlPFC inactivation, as reported here) can be used to select among such models. For instance, we can estimate the weights of the feedback connections between vlPFC and the ventral stream nodes such that the model approximates the neural firing rates at its IT layer (as measured here) upon random (~0.4 cm^3) lesions of the vlPFC module.

Many studies (Ganis et al., 2007; Harth et al., 1987; Tang et al., 2018) propose a cognitive role of the prefrontal feedback, the idea that these recurrent connections carry an expectation signal that augments the representation of object identity in the IT cortex. In a study conducted by Martin et al. (2019), they provided behavioral and electroencephalogram (EEG)-based evidence that rapid top-down feedback from frontal areas, following a feedforward pass, reshapes the bottom-up responses in lower (occipitotemporal) areas. Our results are consistent with these and other similar conceptual theories. However, those ideas are not specific enough to be tested for individual images. That is, they do not specify how to build an accurate image-computable neural network model of the IT-to-PFC-to-IT circuit. While the results presented in this study do not provide a precise blueprint for such a model, the temporal and image-level specificity that they build on is already useful for guiding the development of new recurrent, image-computable models (Kubilius et al., 2019), and the current results can further guide the placement and simulation testing of a vlPFC node in such models (see more below).

Role of vlPFC in Core Object Recognition Behavior

Previous work (Freiwald et al., 2009; Hung et al., 2005; Kar et al., 2019; Logothetis and Sheinberg, 1996; Majaj et al., 2015) has linked neuronal responses in the IT cortex to primate core object recognition behavior. For instance, Majaj et al. (2015) experimentally rejected a large number of alternative models that link ventral stream population activity to core object recognition behavior (aka "decoding models" or "linking models") in support of a simple linear weighted sum of IT response model. These models posit that the mechanisms of core object recognition beyond IT are approximately linear sums of the activity levels of individual IT neurons computed by neurons in PFC, perirhinal cortex, or the caudate. Using various combinations of model parameters (e.g., numbers of neurons, amount of experience with each object category, and brain location of the downstream linear summing), multiple linking hypotheses can be constructed. Our results do not narrow the space of hypotheses to a single linking model. However, these experiments provide two architectural constraints for new models, and our data can be used to falsify or support each such model. First, based on the behavioral deficits observed upon its inactivation, we infer that vlPFC is required to support core object recognition behavior and therefore needs to be integrated into any future model of such behavior. Second, based on both the OST specificity of the behavioral deficits and deterioration of IT decodes, we infer that feedback signals conveyed via the recurrent connections between vlPFC and the ventral stream (most likely the IT cortex) are likely necessary to support this behavior at its normal level of performance. Below, we speculate and discuss candidate linking models that can be further developed and tested using our results.

One possibility is that the downstream summing nodes (as posited by previous studies) are vlPFC neurons and that those vlPFC neurons drive the monkey's behavior. According to this hypothesis, vlPFC is an additional, bidirectionally connected node of processing that intervenes between IT and behavior. This idea is conceptually simple, and it is motivated by previous data from vlPFC, including results showing that category training in monkeys causes PFC neuronal responses to become categorical-like (Freedman et al., 2001), which is what would be expected if vlPFC was the location of those learned sums of IT

neuronal responses described above. This hypothesis predicts that vlPFC inactivation should lead to an equal decrease in behavioral performance for every image. An alternate possibility, however, is that vlPFC neurons do not drive behavior directly but instead transmit the product of their computations to support recurrently connected efferent targets, such as the IT cortex, which then drives behavior via other brain nodes such as caudate. This second possibility is also consistent with the prior work that demonstrated category selectivity in vlPFC neurons (Freedman et al., 2001). Our data do not unequivocally resolve among these two possibilities. Our results—that vlPFC inactivation leads to larger deficits for late-solved images (Figure 5C)—are consistent with the second possibility. However, the fact that vlPFC inactivation also led to lower but significant deficits for early-solved images argues for some element of the first possibility. Indeed, our results overall seem to suggest that both ideas may be partially correct.

Interestingly, Minamimoto et al. (2010) showed that monkeys with bilateral removal of lateral PFC seamlessly learn and generalize perceptual categories. Our data, as presented here, are not necessarily in direct contradiction to these prior results. First, it is essential to note key differences in the two approaches (for a discussion, see Jazayeri and Afraz, 2017), a much smaller, unilateral, reversible inactivation protocol in our case versus a large bilateral permanent lesion in Minamimoto et al. (2010). Second, upon visual inspection, the images used in the Minamimoto et al. (2010) study resemble our early-solved images with canonical views of the objects around the fixation point (without variation in size, position, and other factors) and without any image background (which they explicitly removed in their study). They also showed the images for much longer durations (0.5–1.5 s) along with a pre-cue. Our data confirm that inactivation of vlPFC indeed produces significantly weak behavioral deficits, for early-solved images (see Figure S3B for more details). We do not claim that top-down influence from vlPFC is equally critical for learning and generalization across all object categories and images. Instead, our data suggest that feedback from vlPFC preferentially boosts the IT population code and subsequent behavioral accuracy in specific late-solved images during a rapid categorization task.

Experimentally, we speculate that large-scale neural measurements in brain-regions like vlPFC, collected simultaneously in behaving monkeys (solving a wide variety of recognition tasks), will be required to gain further insights. Furthermore, feedback projection-specific causal perturbation experiments (similar to Oguchi et al., 2015) will be necessary to identify and functionally characterize some of these circuit motifs. Our results suggest that if we are able to specifically inactivate vlPFC to ventral stream recurrent pathways, it will not completely disrupt the core recognition behavior, but it will reduce the primates' performance for certain specific images. To drive further progress, we now also need to incorporate the hypothesized circuit motifs (including a recurrently connected vlPFC node) and build specific artificial neural network models motivated by these experimental results, test their image-by-image predictions (Bashivan et al., 2019; Schrimpf et al., 2018), eliminate models that do not match the experimental data, and build new models. We can use the measured IT neural responses and behavioral accuracies per image (reported in this study; both with and without the inactivation of vlPFC) to constrain as well as validate these new recurrent neural network models of core recognition. That iterative cycle will ultimately lead to a complete, neurally mechanistic understanding of visual object recognition, from images to behavior.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Visual stimuli: generation
  - Generation of synthetic (“naturalistic”) images
  - Generation of natural images (photographs)
  - Primate behavioral testing
  - Large scale multi-electrode recordings and simultaneous pharmacological inactivation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Behavioral Metrics
  - Estimation of neural onset latency per vlPFC site
  - Neural recording quality metrics per IT site
  - Estimation of IT population decode accuracies at OST
  - Estimation and comparison of IT decodes at specific thresholds per image
  - Changes in trial by trial IT decodes upon vlPFC inactivation
  - Estimating change in image-driven IT response rank order (early versus late)
  - Binary object discrimination tasks with DCNNs
  - Prediction of IT neural responses from Deep Convolutional Neural Networks (DCNN) features

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.neuron.2020.09.035.

ACKNOWLEDGMENTS

This research was supported by the Office of Naval Research (grant MURI-114407 to J.J.D) and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We thank K.M. Schmidt, A.R. Murthy, and S. Sanghavi for technical assistance and T. Marques for comments on manuscript.

AUTHOR CONTRIBUTIONS

K.K. and J.J.D. designed the experiments and data analyses pipelines. K.K. carried out the experiments. K.K. performed the data analyses. K.K. and J.J.D. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 14, 2020
Revised: June 5, 2020
Accepted: September 25, 2020
Published: October 19, 2020

REFERENCES

Arikan, R., Blake, N.M., 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.

Bashivan, P., Kar, K., and DiCarlo, J.J. (2019). Neural population control via deep image synthesis. Science *364*, eaav9436.

Borra, E., Ichinohe, N., Sato, T., Tanifuji, M., and Rockland, K.S. (2010). Cortical connections to area TE in monkey: hybrid modular and distributed organization. Cereb. Cortex *20*, 257–270.

Bullier, J., Hupe, J.M., James, A.C., and Girard, P. (2001). The role of feedback connections in shaping the responses of visual cortical neurons. Prog. Brain Res. *134*, 193–204.

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

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science *291*, 312–316.

Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J. Neurosci. *23*, 5235–5246.

Freiwald, W.A., Tsao, D.Y., and Livingstone, M.S. (2009). A face feature space in the macaque temporal lobe. Nat. Neurosci. *12*, 1187–1196.

Fuster, J.M., Bauer, R.H., and Jervey, J.P. (1985). Functional interactions between inferotemporal and prefrontal cortex in a cognitive task. Brain Res. *330*, 299–307.

Ganis, G., Schendan, H.E., and Kosslyn, S.M. (2007). Neuroimaging evidence for object model verification theory: Role of prefrontal control in visual object categorization. Neuroimage *34*, 384–398.

Geirhos, R., Janssen, D.H., Schütt, H.H., Rauber, J., Bethge, M., and Wichmann, F.A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv, 1706.06969 https://arxiv.org/abs/1706.06969.

Gerbella, M., Belmalih, A., Borra, E., Rozzi, S., and Luppino, G. (2010). Cortical connections of the macaque caudal ventrolateral prefrontal areas 45A and 45B. Cereb. Cortex *20*, 141–168.

Harth, E., Unnikrishnan, K.P., and Pandya, A.S. (1987). The inversion of sensory processing by feedback pathways: a model of visual cognitive functions. Science *237*, 184–187.

Hempel, C.G. (1966). Philosophy of Natural Science (Prentice-Hall).

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.

Hupé, J.M., James, A.C., Payne, B.R., Lomber, S.G., Girard, P., and Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. Nature *394*, 784–787.

Hwang, J., and Romanski, L.M. (2015). Prefrontal neuronal responses during audiovisual mnemonic processing. J. Neurosci. *35*, 960–971.

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

Kar, K., Kubilius, J., Schmidt, K., Issa, E.B., and DiCarlo, J.J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat. Neurosci. *22*, 974–983.

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K.A., Cichy, R.M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. USA *116*, 21854–21863.

Kravitz, D.J., Saleem, K.S., Baker, C.I., Ungerleider, L.G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn. Sci. *17*, 26–49.

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., and Schmidt, K. (2019). Brain-like object recognition with high-performing shallow recurrent anns. arXiv, 1909.06161 https://arxiv.org/abs/1909.06161.

Lehky, S.R., and Tanaka, K. (2016). Neural representation for object recognition in inferotemporal cortex. Curr. Opin. Neurobiol. *37*, 23–35.

Liao, Q., and Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv, 1604.03640 https://arxiv.org/abs/1604.03640.

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

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.

Martin, J.G., Cox, P.H., Scholl, C.A., and Riesenhuber, M. (2019). A crash in visual processing: Interference between feedforward and feedback of successive targets limits detection and categorization. J. Vis. *19*, 20.

McKee, J.L., Riesenhuber, M., Miller, E.K., and Freedman, D.J. (2014). Task dependence of visual and category representations in prefrontal and inferior temporal cortices. J. Neurosci. *34*, 16065–16075.

Minamimoto, T., Saunders, R.C., and Richmond, B.J. (2010). Monkeys quickly learn and generalize visual categories without lateral prefrontal cortex. Neuron *66*, 501–507.

Monosov, I.E., Sheinberg, D.L., and Thompson, K.G. (2011). The effects of prefrontal cortex inactivation on object responses of single neurons in the inferotemporal cortex during visual search. J. Neurosci. *31*, 15956–15961.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J.J., and Yamins, D.L. (2018). Task-driven convolutional recurrent models of the visual system. arXiv, 1807.00053 https://arxiv.org/abs/ 1807.00053.

Oguchi, M., Okajima, M., Tanaka, S., Koizumi, M., Kikusui, T., Ichihara, N., Kato, S., Kobayashi, K., and Sakagami, M. (2015). Double virus vector infection to the prefrontal network of the macaque brain. PLoS ONE *10*, e0132825.

Partsalis, A.M., Zhang, Y., and Highstein, S.M. (1995). Dorsal Y group in the squirrel monkey. II. Contribution of the cerebellar flocculus to neuronal responses in normal and adapted animals. J. Neurophysiol. *73*, 632–650.

Popper, K.R. (1959). The Logic of Scientific Discovery (Basic Books).

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.

Riesenhuber, M., and Poggio, T. (2000). Models of object recognition. Nat. Neurosci. *3*, 1199–1204.

Sandell, J.H., and Schiller, P.H. (1982). Effect of cooling area 18 on striate cortex cells in the squirrel monkey. J. Neurophysiol. *48*, 38–48.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., and Schmidt, K. (2018). Brain-score: which artificial neural network for object recognition is most brain-like? bioRxiv. https://doi.org/10.1101/407007.

Spoerer, C.J., McClure, P., and Kriegeskorte, N. (2017). Recurrent convolutional neural networks: a better model of biological object recognition. Front. Psychol. *8*, 1551.

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

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. Proc. Natl. Acad. Sci. USA *115*, 8835–8840.

Tomita, H., Ohbayashi, M., Nakahara, K., Hasegawa, I., and Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature *401*, 699–703.

Wang, C., Waleszczyk, W.J., Burke, W., and Dreher, B. (2000). Modulatory influence of feedback projections from area 21a on neuronal activities in striate cortex of the cat. Cereb. Cortex *10*, 1217–1232.

Webster, M.J., Bachevalier, J., and Ungerleider, L.G. (1994). Connections of inferior temporal areas TEO and TE with parietal and frontal cortex in macaque monkeys. Cereb. Cortex *4*, 470–483.

Yamins, D.L., and DiCarlo, J.J. (2016). Eight open questions in the computational modeling of higher sensory cortex. Curr. Opin. Neurobiol. *37*, 114–120.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. USA *111*, 8619–8624.

Yeterian, E.H., Pandya, D.N., Tomaiuolo, F., and Petrides, M. (2012). The cortical connectivity of the prefrontal cortex in the monkey brain. Cortex *48*, 58–81.

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

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	The Mathworks Inc.	9.8.0.1359463 (R2020a) Update 1

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Kohitij Kar (kohitij@mit.edu).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

At the time of publishing, the behavioral, neural, and modeling data will be available upon reasonable request from the lead contact. In addition, we will also host the images, primate behavioral and neural benchmarks, and the modeling results at http://www.brain-score.org. For additional code to produce DCNN model fits for neural data refer to https://github.com/kohitij-kar/prediction_demo.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The nonhuman primate subjects in our experiments were two adult (Monkey N: age 9 years, and monkey B: age 5 years) male rhesus monkeys (*Macaca mulatta*).

METHOD DETAILS

Visual stimuli: generation

All stimuli used in this study were previously used in the Kar et al. (2019) study and can be accessed at https://github.com/kohitij-kar/image_metrics. For a brief description of the stimuli, please refer below.

Generation of synthetic ("naturalistic") images

High-quality images of single objects were generated using free ray-tracing software (http://www.povray.org), similar to Majaj et al. (2015). Each image consisted of a 2D projection of a 3D model (purchased from Dosch Design and TurboSquid) added to a random background. The ten objects chosen were bear, elephant, face, apple, car, dog, chair, plane, bird and zebra **(**Figure 5A**)**. By varying six viewing parameters, we explored three types of identity while preserving object variation, position (*x* and *y*), rotation (*x*, *y*, and *z*), and size. All images were achromatic with a native resolution of 256 × 256 pixels.

Generation of natural images (photographs)

Images pertaining to the ten nouns, was download from https://cocodataset.org/. Each image was resized to 256 × 256 × 3 pixel size and presented within the central 8°. We used the same images while testing the feedforward DCNNs.

Primate behavioral testing

Active binary object discrimination task

We measured monkey behavior from two male rhesus macaques. Images were presented on a 24-inch LCD monitor (1920 × 1080 at 60 Hz) positioned 42.5 cm in front of the animal. Monkeys were head fixed. Monkeys fixated a white dot (0.2°) for 300 ms to initiate a trial. The trial started with the presentation of a sample image (from a set of 1320 images) for 100 ms. This was followed by a blank gray screen for 100 ms, after which the choice screen was shown containing a standard image of the target object (the correct choice) and a standard image of the distractor object. The monkey was allowed to view the choice objects freely for up to 1500 ms and indi-

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

cated its final choice by holding fixation over the selected object for 400 ms. Trials were aborted if gaze was not held within ± 2° of the central fixation dot during any point until the choice screen was shown We have presented our results after the data was pooled across both monkeys.

Passive Fixation Task

During the passive viewing task, monkeys fixated a white dot (0.2°) for 300 ms to initiate a trial. We then presented a sequence of 5 to 10 images, each ON for 100 ms followed by a 100 ms gray (background) blank screen. This was followed by fluid reward and an inter trial interval of 500 ms, followed by the next sequence. Trials were aborted if the gaze was not held within ± 2° of the central fixation dot during any point.

Data collection

We divided the data collection into two different sessions (days with and without muscimol injections; Figure 3A). These two sessions were repeated in the same order with a minimum gap of one day post the muscimol session (Figure 3B; experimental timeline). On each session (day), monkeys performed the following tasks sequentially: a passive fixation task, a binary object discrimination task, a second passive fixation task. On the second session (day), after the initial passive fixation task (which was included in the no-muscimol condition during all the analyses), we injected a total of 10μl of muscimol at 5 depths (2 μl each) separated by 0.5 mm in the previously localized vlPFC area (for details see below).

Large scale multi-electrode recordings and simultaneous pharmacological inactivation

Surgical implant of chronic micro-electrode arrays

We surgically implanted each monkey with a head post under aseptic conditions. After behavioral training, we recorded neural activity using 10 × 10 micro-electrode arrays (Utah arrays; Blackrock Microsystems). A total of 96 electrodes were connected per array. Each electrode was 1.5 mm long and the distance between adjacent electrodes was 400 μm. Before recording, we implanted each monkey multiple Utah arrays in the IT cortex (monkey B: 2 arrays in left hemisphere); monkey N: 2 arrays in the right hemisphere; shown schematically in Figure 4A). Array placements were guided by the sulcus pattern, which was visible during surgery. The electrodes were accessed through a percutaneous connector that allowed simultaneous recording from all 96 electrodes from each array. Behavioral testing was performed using standard operant conditioning (fluid reward), head stabilization, and real-time video eye tracking. All surgical and animal procedures were performed in accordance with National Institutes of Health guidelines and the Massachusetts Institute of Technology Committee on Animal Care.

Surgical implant of vlPFC injection chamber

During the same surgery, as the chronic array implant, we also placed a semi-cylindrical chamber (Crist Instruments) over a craniotomy targeting the prefrontal cortex, around the principal sulcus. We placed the chambers in the left and right hemispheres of monkey B and monkey N respectively. The chambers were held in place by dental acrylic (methyl methacrylate) applied around the chamber. We used previously reported anatomical landmarks (Freedman et al., 2003; McKee et al., 2014; Tomita et al., 1999), identified by an initial MRI, to guide the vlPFC chamber placements (approximate AP extent shown in Figure 2A). The target injection locations are also consistent with previous reports of visually responsive neurons in vlPFC (Hwang and Romanski, 2015).

vlPFC injection protocol

During the sessions with muscimol injections, we first carefully scraped the dura for maximal visibility and minimum resistance in the path of injection. Then, we used an in-house set up to lower the injection needles (30-32 gauge, small Hub RN Needle; Hamilton Company) using a micro-syringe pump and controller (Micro4^{TM} World Precision Instruments). We started approximately 3 mm below the estimated surface of the dura. We injected 0.5 uL of muscimol (5mg/mL, Sigma Aldrich) at that depth at a speed of 1000 nL/min and, waited for 3 mins and pulled the needle up by ~0.5 mm. This was repeated for 5 depths in total. After the end of the final injection, we waited for 30 mins before the start data collection. Previous works (Arikan et al., 2002; Partsalis et al., 1995) suggest that each of our injections should approximately affect ~2mm diameter of spherical volume of tissue around the injection site.

Eye Tracking

We monitored eye movements using video eye tracking (SR Research EyeLink 1000). Using operant conditioning and water reward, our 2 subjects were trained to fixate a central white dot (0.2°) within a square fixation window that ranged from ± 2°. At the start of each behavioral session, monkeys performed an eye-tracking calibration task by making a saccade to a range of spatial targets and maintaining fixation for 500 ms. Calibration was repeated if drift was noticed over the course of the session.

Electrophysiological Recording

During each recording session, band-pass filtered (0.1 Hz to 10 kHz) neural activity was recorded continuously at a sampling rate of 20 kHz using Intan Recording Controller (Intan Technologies, LLC). The majority of the data presented here were based on multi-unit activity. We detected the multi-unit spikes after the raw data was collected. A multi-unit spike event was defined as the threshold crossing when voltage (falling edge) deviated by more than three times the standard deviation of the raw voltage values. Of 384 implanted electrodes, 2 arrays (left and right hemispheres for monkey B and N respectively) × 96 electrodes × two monkeys, we focused on the 153 most visually driven, and reliable neural sites. Our array placements allowed us to sample neural sites from different parts of IT, along the posterior to anterior axis. However, for all the analyses, we did not consider the specific spatial location of the site, and treated each site as a random sample from a pooled IT population.

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

QUANTIFICATION AND STATISTICAL ANALYSIS

Behavioral Metrics

We have used a one-versus-all image level behavioral performance metric (similar to the one used in Kar et al., 2019) to quantify the behavioral performance of the monkeys as well as DCNNs (described below). This metric estimates the overall discriminability of each image containing a specific target object from all other objects (pooling across all 9 possible distractor choices).

Given an image of object '*i*', and all nine distractor objects ($j \neq i$) we computed the average performance per image as,

$$Performance^{i}_{image} = \frac{\sum_{j=1}^{10} Pc^{i,j\neq i}_{image}}{9},$$

where *Pc* refers to the fraction of correct responses for the binary task between objects '*i*' and '*j*'.

To compute the reliability of this vector, we split the trials per image into two equal halves by resampling without substitution. The median of the Spearman-Brown corrected correlation of the two corresponding vectors (one from each split half), across 1000 repetitions of the resampling was then used as the reliability score (i.e., internal consistency).

Estimation of neural onset latency per vlPFC site

We first normalized (z-scored) the average neural responses (across 80 images and 10 repetitions per image) per site. The onset latencies were then determined as the earliest time from image onset when the firing rates of neurons were higher than one-tenth of the peak of its response. These values have been reported in Figure 2D.

Neural recording quality metrics per IT site

Visual drive per IT neuron (d'_{visual})

We estimated the overall visual drive for each electrode. This metric was estimated by comparing the image responses of each site to a blank (gray screen) response.

$$d'_{visual} = \frac{avg(R_{images}) - avg(R_{gray})}{\sqrt{\frac{1}{2}\left(\sigma^2_{R_{images}} + \sigma^2_{R_{gray}}\right)}}$$

Image rank-order response reliability per neural site (ρ^{IRO}_{site})

To estimate the reliability of the responses per site, we computed a Spearman-Brown corrected, split half (trial-based) correlation between the rank order of the image responses (all images).

Inclusion criterion for IT neural sites

For our analyses, we only included the neural recording sites that had an overall significant visual drive (d'_{visual}), *and an* image rank order response reliability (ρ^{IRO}_{site}) *that was greater than 0.6.* Given that most of our neural metrics are corrected by the estimated noise at each neural site, the criterion for selection of neural sites is not that critical. It was mostly done to reduce computation time and eliminate noisy recordings.

Estimation of IT population decode accuracies at OST

To estimate what information downstream neurons could easily "read" from a given IT neural population, we used a simple, biologically plausible linear decoder (i.e., linear classifiers), that has been previously shown to link IT population activity and primate behavior (Majaj et al., 2015). Such decoders are simple in that they can perform binary classifications by computing weighted sums (each weight is analogous to the strength of synapse) of input features and separate the outputs based on a decision boundary (analogous to a neuron's spiking threshold). Here we have used a support vector machine (SVM) algorithm with linear kernels. The SVM learning model generates a decoder with a decision boundary that is optimized to best separate images of the target object from images of the distractor objects. The optimization is done under a regularization constraint that limits the complexity of the boundary. We used L2 (ridge) regularization, where the objective function for the minimization comprises of an additional term (to reduce model complexity),

$$\text{L2(penalty)} = \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2$$

where β and p are the classifier weights associated with 'p' predictors (neurons). A stochastic gradient descent solver was used to estimate 10 (one for each object) one-versus-all classifiers. After training each of these classifiers with a set of 100 training images per object, we generated a class score (*sc*) per classifier for all held out test images given by,

$$sc = R\beta + bias,$$

where R is the population response vector and the bias is estimated by the SVM solver. The train and test sets were pseudo-randomly chosen multiple times until we every image of our image set was part of the held-out test set. Only the responses from the no-muscimol conditions were treated as training signal. All predictions were made either on held-out responses from no-muscimol or

Please cite this article in press as: Kar and DiCarlo, Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition, Neuron (2020), https://doi.org/10.1016/j.neuron.2020.09.035

muscimol conditions. We then converted the class scores into probabilities by passing them through a *softmax* (normalized exponential) function.

$$P^{i}_{image} = \frac{e^{sc_i}}{\sum_{i=1}^{10} e^{sc_i}}$$

In our previous study (Kar et al., 2019), object solution time per image, OST_{image} was defined as the time it takes for linear IT population decodes to reach within the error margins of the pooled monkey behavioral accuracy for that image. Given that we have used the exact same images in this study, we have used our previously estimated OST per image as the time point of comparison of IT decode accuracy, P^{i}_{image} (with and without muscimol) per image. All reported values of IT population decode accuracies are estimates of how well the population decode accuracy was at the specific OST estimated for the specific image.

Estimation and comparison of IT decodes at specific thresholds per image

We estimated how quickly the IT decodes (based on the pooled IT neurons in the monkeys used in the current study; n = 153 sites) evolve to a specific threshold (mean percent correct values of 0.6) per image. This is another way of approximating which images evolve faster to their identity solutions (in IT) compared to others. The IT decode accuracies per 10 ms time bins were estimated identical to the method discussed above. Based on these estimates, we then divided the images into two groups; ''fast decoded'' (180 images requiring < 120 ms to reach accuracy of 0.6) and ''slow decoded'' (210 images; requiring > 150 ms to reach accuracy of 0.6). These images were also screened such that the overall behavioral performance of the monkey was not significantly different across the two groups. We observed that the ''slow-decoded'' images showed significantly larger deficits (unpaired t test; $t(388) = 3.1$; $p < 0.001$) upon vlPFC inactivation, supporting our primary results.

Changes in trial by trial IT decodes upon vlPFC inactivation

First, we tested how well IT neurons can predict image by image behavioral accuracy patterns (one-versus-all image level behavioral performance metric explained above). To model how downstream neurons might ''read'' IT population responses to infer object identities, we constructed multiple candidate linking models that convert neural responses into a prediction of behavioral choice. Each of the 205 tested linking models was built using neural response data from Kar et al. (2019). We used similar linear SVMs (as mentioned above) with L2 regularization, that differed only in the temporal integration windows (e.g., 70 to 170 ms post image onset, 150 to 200 ms post image onset etc.). We observed that decoders that relied on responses summed from 180 to 220 ms post image onset were most consistent with the pattern of monkey behavioral accuracies (large red dot; Figure S2A). We then used this decoding model to estimate how well such a decoder might predict trial by trial behavioral responses across the two monkeys tested here. Once we estimated the probability of a response per trial, we then performed a ROC analysis to estimate the area under the ROC (AUROC), by taking into account the decoder accuracies for the correct and the incorrect trials. We observed that there was a (slightly) above chance probability of such IT decodes to predict trial by trial monkey choices. However, we found no significant difference between the AUROC while comparing the muscimol and no-muscimol conditions (Figure S2C). For the statistical test, we performed a permutation test by combining data across the two conditions, generating a null distribution of differences in AUROC and comparing it against the true measured difference.

Estimating change in image-driven IT response rank order (early versus late)

For each neuron, we estimated the image response vector ($\vec{r^i}$) at two specific time bins (early: 90-120 ms, and late: 150-180 ms; post image onset). To estimate the change in this vector across time, we computed the noise corrected correlation between the $\vec{r^i}$ vectors estimated at the early and late time bins respectively, as follows

$$\frac{r(early, late)}{\sqrt{\rho(early)} * \sqrt{\rho(y,y)}}$$

Where r(early, late) is the correlation between the $\vec{r^i}$ vectors estimated at the early and late time bins, and ρ(early) and ρ(late) are the split-half (across trial) reliability of these vectors estimated independently at the corresponding time bins. We computed these noise-corrected correlation values per neuron for both the no-muscimol and muscimol conditions.

Binary object discrimination tasks with DCNNs

We have used the same linear decoding scheme mentioned above (for the IT neurons) to estimate the object identity solution strengths per image for the DCNNs. Briefly, we first obtained an imagenet pre-trained DCNN (e.g., AlexNet). We then replaced the last three layers (i.e., anything beyond 'fc7') of this network with a fully connected layer containing 10 nodes (each representing one of the 10 objects we have used in this study). We then trained this last layer with a back-end classifier (L2 regularized linear classifier; similar to the one mentioned for IT) on a subset of images from our image-set. These images were selected randomly from our imageset and used as the train-set. The remaining images were then used for the testing (such that there is no overlap between the train and test images). Repeating this procedure multiple times allowed us to use all images as test images providing us with the performance of the model for each image.

To compute the behavioral predictivity score, we correlated (Pearson correlation) the averaged performance estimated per image across the pooled monkey data and that estimated from the DCNN. The correlation scores were further corrected by the trial-level split half reliability of the monkey data and DCNN behavioral scores (split half reliability for DCNN = 1, since estimates are noiseless).

Prediction of IT neural responses from Deep Convolutional Neural Networks (DCNN) features

We modeled each IT neural site as a linear combination of the DCNN model features. We first extracted the features per image, from the DCNNs' penultimate layers. Using a 10-fold train/test split of the images, we then estimated the regression weights (i.e., how we can linearly combine the model features to predict the neural site's responses) using a partial least square (MATLAB command: *plsregress*) regression procedure, using 20 retained components. For each set of regression weights estimated on a train imageset, we generated the output of that 'synthetic neuron' for the held-out test set. The percentage of explained variance, *IT predictivity* (for more details refer Yamins et al., 2014) for that neural site, was then computed by normalizing the r^2 prediction value for that site by the self-consistency of the image responses for that site and the self-consistency of the regression weights for that site (estimated by a Spearman Brown corrected trial-split correlation score). Table S1 lists all the models we have tested and the corresponding evaluated layers.