Evidence that the ventral stream codes the errors used in hierarchical inference and learning

Running title: Error coding in the ventral stream

Elias B. Issa*, Charles F. Cadieu, and James J. DiCarlo

McGovern Institute for Brain Research and Department of Brain and Cognitive Sciences
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

*Elias Issa, McGovern Institute for Brain Research, Department of Brain and Cognitive Sciences, 77 Massachusetts Institute of Technology, 46-6157, Cambridge, MA 02139. E-mail: issa@mit.edu

AUTHOR CONTRIBUTIONS
E.B.I. and J.J.D designed the experiments. E.B.I. carried out the experiments and performed the data analysis. E.B.I. and C.F.C. designed the models. E.B.I., C.F.C., and J.J.D. wrote the manuscript.

ABSTRACT

Hierarchical feedforward processing makes object identity explicit at the highest stages of the ventral visual stream. We leveraged this computational goal to study the fine-scale temporal dynamics of neural populations in posterior and anterior inferior temporal cortex (pIT, aIT) during face detection. As expected, we found that a neural spiking preference for natural over distorted face images was rapidly produced, first in pIT and then in aIT. Strikingly, in the next 30 milliseconds of processing, this pattern of selectivity in pIT completely reversed, while selectivity in aIT remained unchanged. Although these dynamics were difficult to explain from a pure feedforward perspective, a model class computing errors through feedback closely matched the observed neural dynamics and parsimoniously explained a range of seemingly disparate IT neural response phenomena. This new perspective augments the standard model of online vision by suggesting that neural signals of states (e.g. likelihood of a face being present) are intermixed with the error signals found in deep hierarchical networks.

INTRODUCTION

The primate ventral visual stream is a hierarchically organized set of cortical areas beginning with the primary visual cortex (V1) and culminating with explicit (i.e. linearly decodable) representations of objects in inferior temporal cortex (IT) (1) that quantitatively account for invariant object discrimination behavior (2). Consistent with a feedforward flow of processing from V1 to V2 to V4 to IT, neurons at higher cortical stages are more selective for object shape and identity while being more tolerant to changes in object size and position (3)(4). Formalizing object recognition as the result of a series of feedforward computations yields models that achieve impressive performance on basic object categorization tasks (5)(6) similar to the level of performance achieved by IT neural populations (7)(8). Importantly, these models are only optimized to solve invariant object recognition tasks and are not trained on neural data, yet their intermediate layers are highly predictive of time-averaged V4 and IT neural responses supporting the longstanding belief that the computational purpose of the ventral stream is to solve such tasks. Thus, the feedforward inference perspective provides a simple but powerful, first-order framework for studying core invariant object recognition (9).

However, visual object recognition behavior may not simply be the result of a single feedforward neural processing pass (a.k.a. feedforward inference), as evidenced by the

observed dynamics of neural responses in IT (10)(11)(12)(13). From a theoretical perspective, feedback in a recurrent network might substantially improve online inference or support learning (14)(15). For example, inferring that a particular object is present in an image in the presence of local ambiguities (e.g. ambiguity due to noise, missing local information, or local information that is incongruent with the whole) can be achieved by integrating top-down with bottom-up information through feedback (16)(17)(18). Another possibility is that feedback from high level representations to lower level representations is used to drive changes of the synaptic weights of neural networks (i.e. learning) through unsupervised or supervised algorithms (e.g. error backpropagation) so as to improve future feedforward inference in performing tasks such as object recognition (19). These two possibilities are not mutually exclusive, and several computational models aim to integrate both in a natural way by building feedback pathways onto a feedforward architecture (see **Fig. 7** and **Discussion**).

These architectures constitute a broad class of recurrent networks that all have an error computation which compares top-down predictions or targeted outputs against bottom-up inputs to compute gradients for driving sensory inference and learning (19)(20). Though the need for errors and subsequent gradient signals is theoretically well motivated, we do not know if such computations (e.g. error backpropagation and hierarchical Bayesian inference) are implemented in the brain. Currently, little conclusive neurophysiological evidence exists in support of these models. For example, while it has been suggested that end-stopped tuning of V1 cells is the result of feedback loops that “explain away” extended edges (21), this feedback-based interpretation cannot be disambiguated from standard alternative interpretations utilizing local mechanisms like lateral inhibition (which builds selectivity for shorter edges) (22) or normalization (which reduces responses to large stimuli such as extended edges) (23). Thus, these neural measurements taken from a single cortical area in which the mapping to a visual behavior is unclear, only weakly constrain computational models, and an open question is whether the dynamics of propagation of neural signals across stages of the visual hierarchy performs anything beyond feedforward inference.

Here, to disambiguate between the different types of inference that can be implemented in a hierarchy through feedforward, lateral, or feedback connections, we measured the temporal dynamics of neural signals across three stages of face processing in macaque IT. We found that many neurons in the intermediate processing stages reversed their initial preference – they rapidly switched from face preferring to anti-face preferring. Standard feedforward models including those employing local recurrences such as adaptation, lateral inhibition, and normalization could not fully capture the dynamics of face selectivity in our data. Instead, our modeling revealed that the reversals of face selectivity in intermediate processing stages are a natural dynamical signature of a family of hierarchical models that use feedback connections to implement “error coding.” We interpret this very good fit to our data as evidence that the ventral stream is implementing a model in this family. If correct, this model family informs us that we should not interpret neural spiking activity at each level of the ventral stream as only an explicit representation of the variables of interest (e.g. is a face present?) but should interpret much of spiking activity as representing the necessary layer-wise errors propagated across the hierarchy. In addition, we find that, without additional parameter modifications, the same error coding hierarchical model family explains seemingly disparate IT neural response phenomena, hence unifying our results with previous findings under a single computational framework.

RESULTS

Neural recordings were made across the posterior to anterior extent of IT in the left hemispheres of two monkeys (**Fig. 1**; example neurophysiological map in one monkey). Recording locations were accurately localized *in vivo* and co-registered across penetrations using a stereo microfocal x-ray system (~400 micron *in vivo* resolution)(24). The face versus non-face object selectivity of each site was measured using a standard screen image set, and neural maps of face versus non-face object selectivity were used to physiologically define subregions of IT containing an enrichment of face preferring sites (a.k.a. face patches) (25). Sites were assigned as belonging to a face-preferring cluster by their distance to the center, a purely spatial rather than functional criterion. The identified subregions were present in both monkeys and encompassed at least three hierarchical stages of face processing in posterior (pIT), central (cIT), and anterior (aIT) IT (**Fig. 1**). We asked how neural signals are transformed from the earliest (pIT) to the latest stage (aIT) of IT face processing.

Images with typical versus atypical arrangements of face parts

Converging lines of evidence from correlative and causal studies implicate the face patch subnetwork in face detection (25)(26). However, these studies relied on images testing general face versus non-face object discrimination (a.k.a. face detection) that can be solved using differences across multiple feature dimensions (local contrast, spatial frequency, and number of features). We sought images that would provide a more stringent test of the face subnetwork near the limits of its detection abilities. Specifically, if these regions are performing accurate face detection, they should specifically respond to the configuration of a face even when all other features are matched. Thus, we probed the face selectivity of IT with face-like images that only differed in the configuration of the face parts such that there could be conflicting local (parts) and global (configuration) evidence as to whether a face was truly present (**Fig. 2a**). For example, images with an eye presented in an unnatural, centered position have a very different global interpretation ("cyclops") than when the eye is presented in its natural lateralized position in the outline even though these images have very similar local information (eye presented in an outline in the upper visual field) (see examples in **Fig. 2a** and **Supplementary Fig. 1b**). These images create an "aperture problem" because they are difficult to distinguish based on local information alone and must be disambiguated based on the surrounding context (27). This image set thus poses a more stringent challenge of face detection ability than standard screen sets which vary along many stimulus dimensions (i.e. faces vs bodies and non-face objects; see **Fig. 1**). Consistent with previous work showing that neurons in pIT can be driven by images like the 'cyclops' which contain information that is globally inconsistent with a face (28), we identified 13 atypical face part configurations that drove neurons to produce an early response that was $\geq$90% of their response to a correctly configured whole face (**Fig. 2a**). Because these images drove a high feedforward response, we view them as being relatively well matched in their low-level image properties and capable of activating additional processing in face responsive regions of IT. These images were comparable to whole faces from our screen set in driving responses with short latency (median latency: whole face=70.0 $\pm$ 0.4 ms, typical arrangement of parts=69.5 $\pm$ 0.5 ms, atypical arrangements of parts=70.0 $\pm$ 0.5 ms, n_{images}=10, 8, and 13, respectively; $p < 0.01$ for atypical versus typical and atypical versus whole face, n=115 sites), and firing rates were comparable in strength to whole faces even over a broad analysis window (median normalized response over 50-200 ms: whole face exemplars=1.22 $\pm$ 0.23, typical arrangement=1.36 $\pm$ 0.04, atypical arrangements=1.37 $\pm$ 0.03; $p > 0.05$ for all two-way comparisons, n=115 sites). Thus, building on the insight from previous work that pIT does not perform full face detection and is susceptible to the "aperture problem" even though it passes the basic face versus non-face object test, we constructed images that allowed us to ask how a difficult face detection task might be solved over time in pIT or if a solution is produced in

downstream areas.

Time course of responses in posterior, central, and anterior IT for images with typical versus atypical arrangements of face parts

When we examined the dynamics of neural responses in pIT, we found that an unexpected preference for atypical arrangements of face parts emerged over time (see example sites in **Fig. 2b**). Of the sites showing a significant change in their preference over time, a majority showed a decreasing preference over time for typical arrangements of face parts (43 of 51 sites or 84%; $p < 0.01$ criterion for significant change at the site level, n = 115 sites) (**Fig. 2c**). This surprising trend -- decreasing responses for images with typical as compared to atypical arrangements of face parts -- was strong enough that it reversed the small, initial preference for images of normally arranged face parts over the population (median d': 60-90 ms = 0.11 $\pm$ 0.02 vs. 100-130 ms = -0.12 $\pm$ 0.03, $p < 0.01$, n=115 sites), and this trend was observed in both monkeys when analyzed separately ($p < 0.01$, one-tailed test for decreased d' between 60-90 ms and 100-130 ms in both monkeys, $n_{M1} = 43$, $n_{M2} = 72$ sites; **Fig. 3a**). Even at the individual site level, a complete and rapid reversal of the expected selectivity profile of face neurons could be observed (33 of 51 sites changed from face preferring to non-face preferring and only 3 of 51 sites changed to face preferring after being non-face preferring) (**Fig. 4a**, left). As a result, the majority of pIT sites responded more strongly to atypical images over typical images in the late response phase (prefer typical arrangement: 60-90 ms = 66% vs. 100-130 ms = 34%; $p < 0.01$, n = 115) (**Fig. 4b**, light green bars).

In the anterior face-selective regions of IT which are furthest downstream of pIT and reflect additional stages of feedforward processing (see block diagram in **Fig. 1**), we did not observe these strong reversals in selectivity profile -- 98% of sites (39 of 40) did not change their relative preference for typical vs. atypical face features ($p<0.01$ criterion for significant change at the site level). Rather, we observed a stable selectivity profile over time in aIT (median d': 60-90 ms = 0.13 $\pm$ 0.03 vs. 100-130 ms = 0.17 $\pm$ 0.03, $p > 0.05$, n=40 sites) with a slight but gradual accumulation (increasing preference for normal arrangements of the face parts) rather than a reversal of preference (**Fig. 4a**, right). As a result, the majority of anterior sites preferred images with typical arrangement of the face parts in the late phase of the response (prefer typical: 60-90 ms = 78% of sites vs. 100-130 ms = 78% of sites; $p > 0.05$, n = 40 sites) despite only a minority of upstream sites in pIT preferring these images in their late response (**Fig. 4b**, black bars). This suggests that spiking responses of individual aIT sites resolve images as expected from a computational system whose purpose is to detect faces, as previously suggested (29). Finally, in cIT whose anatomical location is intermediate to PIT and AIT, we observed a selectivity profile over time that was intermediate to that of pIT and aIT consistent with its position in the ventral visual hierarchy (Fig. 4a,b, dark green). The very different patterns of dynamics in pIT, cIT, and aIT neurons are surprising given the intuition that in a feedforward network selectivity for the preferred stimulus class is maintained or built contrary to the complete reversal of rank-order selectivity observed in the lower stages but not the highest stage of IT.

Controls for low-level image variation and for overall activity

To validate our findings against potential confounding factors, we checked the robustness of our surprising observation that preference for typical over atypical arrangements of face parts reverses in pIT. Since the number of parts varied between the two image classes tested, we recomputed selectivity using only images with the same number of parts thus limiting differences in the contrast, spatial frequency and retinal position of energy in the image (see

examples in **Fig. 3b**). We found that pIT face selectivity was still consistently reversed across subsets of images containing a matched number of one, two, or four parts (n=5, 30, and 3 images, respectively). Here, we quantify reversals as a decrease in d' from the early (60-90 ms post-image onset) to the late (100-130 ms post-image onset) response phase where a negative d' indicates a preference for atypical face part arrangements (see **SI Methods**). For all three image subsets controlling the number of face parts to be one, two, or four, d' began positive on average (i.e. preferring typical face part arrangements) (median d' for 60-90 ms = 0.13 $\pm$ 0.05, 0.05 $\pm$ 0.02, 0.33 $\pm$ 0.09 for one, two, and four parts) and significantly decreased in the next phase of the response becoming negative on average (median d' for 100-130 ms: -0.27 $\pm$ 0.06, -0.14 $\pm$ 0.02, -0.04 $\pm$ 0.12; $p < 0.01$ for d' comparisons between 60-90 ms and 100-130 ms, n = 115, 115, 76 sites) (**Fig. 3b**,). A similar reversal in face selectivity was observed when we re-tested single part images at smaller (3°) and larger (12°) image sizes suggesting a dependence on the relative configuration of the parts and not on their absolute retinal location or absolute retinal size (median d' for 60-90 ms vs. 100-130 ms: three degrees = 0.51 $\pm$ 0.09 vs. -0.29 $\pm$ 0.14, twelve degrees = 0.07 $\pm$ 0.14 vs. -0.11 $\pm$ 0.14; n = 15; $p < 0.01$ for three degree condition only) (**Fig. 3c**).

We also tested the hypothesis that absolute initial firing rates, which were not perfectly matched, were somehow responsible for producing the pIT image preference reversal. We found no support for this hypothesis -- the observed change in firing rate over time ($\Delta_{pIT}=r_{pIT\ late} - r_{pIT\ early}$) was weakly correlated with the strength of the initial response ($\rho_{pIT\ early,\Delta pIT} = -0.24 \pm 0.15$, $p = 0.044$, n = 20 images; for these firing rate controls, the original whole face image drove a much higher response than the synthetic images we created, and being a firing rate outlier, we excluded this image). Instead, firing rate changes over time were strongly correlated with the class (typical versus atypical) of the image ($\rho_{class,\Delta pIT} = -0.77 \pm 0.04$, $p < 0.01$, n = 20 images). In other words, responses to images with normally arranged face parts were specifically weaker by 18% on average in the next phase of the response (Δrate (60-90 vs 100-130 ms) = -18% $\pm$ 4%, $p < 0.01$; n = 7 images), but responses to images with unnatural arrangements of face parts, which also drove high initial responses, did not experience any firing rate reduction in the next phase of the response (Δrate (60-90 vs 100-130 ms) = 2 $\pm$ 1%, $p > 0.05$; n = 13 images). This dependence on the image class and not on initial response strength argues against explanations such as rate-driven adaptation that solely depend on a unit's activity to explain decreasing neural responses over time. Indeed, we found that the late phase firing rates in PIT could not be predicted from early phase pIT firing rates ($\rho_{pIT\ early,\ pIT\ late} = 0.07 \pm 0.17$, $p > 0.05$; n = 20 images). In contrast, we found that pIT late phase firing rates were better predicted by early phase firing rates in downstream regions cIT and aIT ($\rho_{cIT\ early,\ pIT\ late} = -0.52 \pm 0.11$, $p < 0.01$; $\rho_{aIT\ early,\ pIT\ late} = -0.36 \pm 0.14$, $p = 0.012$; $n_{pIT}=115$, $n_{cIT}=70$, $n_{aIT}=40$ sites). That is, for images that produced high early phase responses in cIT and aIT, the following later phase responses of units in the lower level area (pIT) tended to be low, consistent with the hypothesis that feedback from those areas is producing the pIT selectivity reversals. Finally, the relative speed of selectivity reversals in pIT (~30 ms peak-to-peak) makes explanations based on fixational eye movements or shifts in attention (e.g. from behavioral surprise to unnatural arrangements of face parts) unlikely as saccades and attention shifts occur on slower timescales (hundreds of milliseconds) (30).

Computational models of neural dynamics in IT

Given the above observations of a non-trivial, dynamic selectivity reversal during face detection, we next proceeded to build formal models of gradually increasing complexity to determine the

minimal set of assumptions that could capture our empirical findings. We used a linear dynamical systems modeling framework to evaluate dynamics in different hierarchical architectures (**Supplementary Fig. 1 a,b** and see **SI Methods**). A core principle of feedforward ventral stream models is that object selectivity is built by feature integration from one cortical area to the next cortical area in the hierarchy leading to low dimensional representations at the top of the hierarchy. Here, we take the simplest feature integration architecture where a unit in a downstream area linearly sums the input from units in an upstream area to produce greater downstream selectivity than any upstream input alone. This generic encoding model conceptualizes the idea that different types of evidence, local (i.e. parts) and global (i.e. arrangement of parts), have to converge and be integrated to separate face from non-face images in our image set. Dimensionality reduction as performed in this network is a key computation specified by most network architectures whether unsupervised (i.e. autoencoder) or supervised (i.e. backprop) allowing these networks to learn an abstracted, low dimensional representation from the high-dimensional input layer. Here, we performed dimensionality reduction in linear networks as monotonic nonlinearities can be readily accommodated in our framework (14)(21). First, we focused on two stage models to use the simplest configuration possible and gain intuition since stacks of two processing stages can be used to generate a hierarchical system of any depth. In this framing, the activity of the unit in the output stage corresponds to aIT which integrates activity from units in the deepest hidden stage measured corresponding to pIT (**Figure 4**, top row, first five models, and **Supplementary Fig. 1 a,b**) bearing in mind that aIT is actually a hidden stage of processing with respect to the next processing stage in the larger cortical stack. When an external step input is applied to such a system, it will of course produce a (lagging) step response in each of the two stages. We here sought to determine how adding recurrent connectivity to this basic feedforward architecture could generate internal dynamics beyond those simple dynamics, and to compare those dynamics with the observed IT neural dynamics.

Based on previous ideas in the literature, we considered lateral inhibition within a stage, normalization within a stage (23), and cortico-cortical feedback (14). Adding recurrent lateral inhibitory connections leads to competition within a stage which can limit responses over time to a strong driving stimulus but not to weaker stimuli. Similarly, normalization scales down responses over time to strong driving stimuli and can be implemented by a leak term that scales adaptation (the degree of decay in the response) and is controlled recursively by the summed activity of the network (23). Besides lateral connections and normalization within a processing stage, feedback connections between stages are a remaining available mechanism for driving dynamics. To constrain our choice of a feedback-based model, we took a normative approach minimizing a quadratic reconstruction cost between stages. While there are many possible layer-wise cost functions to consider (i.e. input to output mappings to optimize), we minimized a classical reconstruction cost as this term is at the core of an array of hierarchical generative models including hierarchical Bayesian inference (31), Boltzmann machines (32), analysis-by-synthesis networks (14), sparse coding (20), predictive coding (21), and autoencoders in general (33). Optimizing a quadratic loss results in feedforward and feedback connections that are symmetric -- reducing the number of free parameters -- such that inference on the represented variables at any intermediate stage is influenced by both bottom-up sensory evidence and current top-down interpretations. Critically, a common feature of this large model family is the computation of between-stage error signals via feedback which is distinct from state-estimating model classes (i.e. feedforward models) that do not compute or propagate errors (see **Figure 7** for a comparison of state and error estimating model classes). A dynamical implementation of such a network uses leaky integration of error signals which, as shared

computational intermediates, guide gradient descent of the values of the represented variables (Δactivity of each neuron => inference) to a previously learned target value or descend the connection weights (Δsynaptic strengths => learning) to values that give the best behavior, here defined as a reconstruction goal (similar results were found using other goals and networks; **Supplementary Fig. 2**).

When we fit each of the models to our neural data, they were all able to produce an increase in selectivity from the first stage of the network to the second stage of the network. This increase is not surprising because all models had converging feedforward connections from the first to second stages (**Figure 5a**, first five columns, compare green and black curves). However, we found that neither the lateral inhibition model, nor the normalization model, could capture the observed selectivity reversal phenomenon in pIT. Instead, the selectivity of these models simply increased to a saturation level set by the leak term (shunting inhibition) in the system (**Figure 5a**, first five columns). Similar behavior was present when we tried a nonlinear implementation of the normalization model that more powerfully modulated shunting inhibition (23). That the normalization models performed poorly can be explained by the fact that responses to a strong stimulus even when normalized can meet but not fall below those to a stimulus that was initially weak. Thus, a complete reversal in stimulus preference on average across a population of cells (i.e. Figures 2-4, PIT data) is not possible when only using normalization mediated by surround suppression.

In contrast to the above models, we found that the feedback model capable of computing hierarchical error signals naturally displayed a strong reversal of selectivity in a sub-component of its first processing stage -- qualitatively similar behavior to the selectivity reversal that we observed in many pIT neural sites. Specifically, this model showed reversal dynamics in the magnitude of its reconstruction error signals but not in its state signals (the states of the world inferred by the model) (**Figure 5a,** compare fifth and sixth columns). These error signals integrate converging state signals from two stages -- one above and one below (see **SI Methods**). The term “error” is thus meaningful in the hidden processing stages where state signals from two stages can converge. The top nodes of a hierarchy receive little descending input and hence do not carry additional errors with respect to the desired computation; rather, top node errors closely approximate the feedforward-driven state estimates (put another way, top stages drive the formation of errors below). This behavior in the higher processing stages is consistent with our observation of explicit representation of faces in aIT in all phases of the response (**Fig. 4**) and with similar observations of decodable identity signals by others in all phases of aIT responses for faces (34) and objects (1)(2).

Finally, we asked whether our results generalized to larger networks of increasing depth. We found similar results for three layer versions of the models described above. Specifically, the dynamics of error signals in a three-layer model produced a good match to our data collected from three successive cortical areas (**Fig. 5a,** seventh column), while state signals in three layer model networks did not produce the observed IT face selectivity dynamics (**Fig. 5b,** right set of bars).

Predictions of an error coding hierarchical model

While our neural observations at multiple stages of the IT hierarchy led us to the error coding hierarchical model above, a stronger test of the idea of error coding is whether it predicts other IT neural phenomena. To identify stimulus regimes that would lead to insightful predictions, we asked in what way would the behavior of error-estimating hierarchical models differ most from

the behavior of generic feedforward state-estimating models. Because our feedback-based model uses feedforward inference at its core, it behaves similarly to a state-estimating hierarchical feedforward model when the statistics of inputs match the learned feedforward weight pattern of the network (i.e. 'natural' images drawn from everyday objects and scenes) since in these settings feedforward inferences derived from the sensory data are aligned with top-down expectations. Thus, predictions of feedback-based models that could distinguish them from feedforward-only models are produced when the natural statistics of images are altered so that they differ from the feedforward patterns previously learned by the network. We have (above) considered one such type of alteration: images where local face features are present but altered from their naturally occurring (i.e. statistically most likely) arrangement. Next, we tested two other image manipulations from recent physiology studies which yielded novel neural phenomena that lacked a principled, model-based explanation (35)(13). To test whether the error coding hierarchical model family displays these behaviors, we fixed the architectural parameters derived from our fitting procedure in **Figure 5** and simply varied the input to this network, specifically the correlation between inputs and the network's weight pattern, in order to match the nature of the image manipulations performed in prior experiments.

Sublinear integration of the face features. In face-selective areas in pIT and cIT, the sum of the responses to the face parts exceeds the response to the whole face (35)(28), and this behavior increases from linear to more sublinear over the timecourse of the response (ratio of sum of responses to parts vs. response to whole: 60-90 ms = 1.5 $\pm$ 0.1, 100-130 ms = 4.6 $\pm$ 0.3; $p < 0.01$, n = 33 sites) (**Fig. 6a**, left panel). This result runs counter to what would be expected in a model where selectivity for the whole face is built from the conjunction of the parts. In such a model, the response to the whole face would be at least as large if not greater (superlinear) than the summed responses to the individual features. To test whether an error coding model exhibited the phenomenon of sublinear feature integration, we compared the response with all inputs active (co-occurring features) to the sum of the responses when each input was activated independently (individual features). The reconstruction errors in our feedback-based model showed a strong degree of sublinear integration of the inputs such that the response to the simultaneous inputs (whole) was much smaller than what would be predicted by a linear sum of the responses to each input alone (parts), and the model's sublinear integration behavior qualitatively replicated the time course observed in pIT without any additional fitting of parameters (**Fig. 6a**, right panel).

Evolution of neural signals across time. Neural responses to familiar images are known to rapidly attenuate in IT when compared to responses to novel images (36)(37)(38). This observation seems to contradict what would be predicted by simple Hebbian potentiation for more exposed stimuli. Furthermore, familiar images show much sharper temporal dynamics than responses to novel images when presented repeatedly (13). These qualitatively different dynamics for familiar versus novel images are surprising given that stimuli are drawn from the same distribution of natural images and are thus matched in their image-level statistical properties (color, spatial frequency, contrast). To test whether our network displayed these different dynamical behaviors, we simulated familiar inputs as those that match the learned weight pattern of a high-level detector and novel inputs as those with the same overall input level but with weak correlation to the learned network weights (here, we have extended the network to include two units in the output stage corresponding to storage of the two familiarized input patterns to be alternated; conceptually, we consider these familiar pattern detectors as existing downstream of IT in a region such as perirhinal cortex which has been shown to code familiarized image statistics and memory-based object signals (39)). We repeatedly alternated

two familiar inputs or two novel inputs and found that model responses in the hidden processing stage were temporally sharper for familiar inputs that matched the learned weight patterns compared to novel, unlearned patterns of input, consistent with the previously observed phenomenon (**Fig. 6b**; data reproduced with permission from Meyer et al., 2014 (13)). Model responses reproduced additional details of the neural dynamics including a large initial peak followed by smaller peaks for responses to novel inputs and a phase delay in the oscillations of responses to novel inputs compared to familiar inputs. Intuitively, these dynamics are composed of two phases. After the initial response transient, familiar patterns lead to lower errors and hence lower neural responses than random patterns (see **Fig. 6b**, red curve drops below the blue curve after the onset response), similar to the observed weaker response to more familiar face-like images present in our data (**Fig. 2d**). When the familiar pattern *A* is switched to another familiar pattern *B*, this induces a short-term error in adjusting to the new pattern (**Fig. 6b**, red curve briefly goes above the blue curve during pattern switch and then decreases). Because unfamiliar patterns are closer together in the high-level encoding space than two learned patterns (**Fig. 6b**, top right panel), the switch between two learned patterns introduces more shift in top-down signals and hence greater change in error signals. This result demonstrates that our model, derived from fitting only the first 70 ms (60-130 ms post image onset) of IT responses to face images, can extend to much longer timescales and may generalize to studies of images besides face images.

Dynamical properties of neurons across cortical lamina

In the large family of state-error coding hierarchical networks, a number of different cortical circuits are possible (**Fig. 7**). A key distinction of two such circuit mapping hypotheses (predictive coding versus error backpropagation) is the expected laminar location of state coding neurons. Specifically, predictive coding asserts that superficial layers contain error units and that errors are projected forward to the next cortical level (21)(40)(41) as opposed to typical neural network implementations where the feedforward projecting neurons in superficial lamina are presumed to encode estimates about states of the visual world (e.g. is a face present or not?). State and error signals can be distinguished by their dynamical signatures in our leading model, which was fit on error signals but produces predictions of the corresponding state signals underlying the generation of errors. Since state units are integrators (see **SI Methods**), they have slower dynamics than error units leading to longer response latencies and a milder decay in responses (**Fig. 6a**, right panel). To test this prediction, we localized our recordings relative to the cortical mantle by co-registering the x-ray determined locations of our electrode (~400 micron *in vivo* accuracy) to structural MRI data (see **SI Methods**). When we separated units into those at superficial depths closer to the pial surface (1/3 of our sites; corresponds to approximately 0 to 1 mm in depth) versus those in the deeper layers (remaining 2/3 of sites, ~1 to 2.5 mm in depth), we found a longer latency and less response decay in superficial units consistent with the expected profile of state units (**Fig. 6a**, left panel). Importantly, the latency difference between cortical lamina within pIT (deep vs superficial: 66.0 ± 1.7 vs 76.0 ± 1.7 ms, $p < 0.01$) was greater than the conduction delay from pIT to cIT (i.e. from superficial layers of pIT to the deeper layers of cIT) (superficial pIT vs deep cIT: 76.0 ± 1.7 vs 75.5 ± 2.1 ms, $p > 0.05$) even though the distance traveled between cortical stages pIT and cIT is larger than laminar distances within pIT. Thus, instead of a simple conduction delay accounting for latency differences across lamina, our model suggests that temporal integration of inputs, consistent with the behavior of state units implemented in standard feature coding rather than predictive coding schemes, may drive the lagged dynamical properties of neurons in superficial lamina.

DISCUSSION

We have measured neural responses during a difficult face detection task across the IT hierarchy and demonstrated that the initial feedforward preference for faces in the intermediate (a.k.a hidden) processing stages reverses over time – that is, after the initial wave of face-selective neural responses, responses at lower levels of the hierarchy (pIT and cIT) rapidly evolve to not prefer typical face part arrangements. This behavior was inconsistent with a pure feedforward model, even when we included strong nonlinearities in these models, such as normalization. However, we showed that augmenting the feedforward model so that it represents the errors generated during hierarchical processing produced the observed neural dynamics (**Fig. 5**). This view argues that a fraction of cortical neurons codes error signals. Using this new modeling perspective, we went on to generate a series of predictions consistent with observed IT neural phenomena (**Fig. 6**). Importantly, this perspective provides an alternative interpretation to prior suggestions that IT neurons are not tuned for typical faces but are instead tuned for atypical faces (35). Under the present hypothesis, some IT neurons are preferentially tuned to typical arrangements of face features, and many other IT neurons are involved in coding errors with respect to those typical arrangements. We believe that these intermixed state estimating and error coding neuron populations are both sampled in standard neural recordings of IT, even though only state estimating neurons are truly reflective of the tuning preferences of that IT processing stage.

The precise fractional contribution of errors to neural activity is difficult to estimate from our data. Under the primary image condition tested, not all sites significantly decreased their selectivity (~60%). We currently interpret these sites as coding state (feature) estimates (**Fig. 2c**, gray and black dots), and we did observe evidence of emergence of state-like signals in our superficial neural recordings (**Fig. 6c**). Alternatively, at least some of the non-reversing sites might be found to code errors under other image conditions than the one that we tested. Furthermore, while in our primary image condition selectivity reversals only accounted for 20% of the overall spiking modulation (**Fig. 2d**), we found larger modulations in late phase neural firing (50-100%) under other image conditions tested (**Fig. 6 a,b**). At a computational level, the absolute contribution of error signals to spiking may not be the critical factor as even a small relative contribution may have important consequences in the network.

Error signals generated across different hierarchical inference and learning models

The notion of error is inherent to many existing models in the literature that go beyond the basic feedforward, feature estimation class. These models use errors for guiding top-down inference by computing errors implicitly (hierarchical Bayesian inference (14)(31); **Fig. 7**, second row) or by representing errors explicitly (predictive coding (21); **Fig. 7**, third and last row). Alternatively, errors can be used specifically for unsupervised learning (autoencoder (33); **Fig. 7**, fourth row) or specifically for supervised learning (classic error backpropagation (19); **Fig. 7**, fifth row). Finally, recent models incorporate aspects of both inference and learning (42)(43) (**Fig. 7**; bottom two rows). A key, unifying feature across inference and learning models is the need to compute an error signal between processing stages. This error signal can be in the form of a generative, reconstruction cost (stage *n* predicting stage *n-1*) or a discriminative, construction cost (stage *n-1* predicting stage *n*). Regardless, this across-stage "performance" error term is used in all models, is typically the only term combining signals from different model layers, and is distinct from within-stage "regularization" terms (i.e. sparseness or weight decay) in driving network behavior. The present study provides evidence that such errors are not only computed, but that they are explicitly encoded in spiking rates. To test the robustness of this claim across

different model implementations, we tested models with different performance errors (reconstruction, nonlinear reconstruction, and discriminative) and found similar population level error signals across these networks (**Supplementary Fig. 2**). Thus, errors as broadly construed in the state-error coding hierarchical model family provide a good approximation to IT population neural dynamics, and future work examining the specifics of error signals and interactions between error and state units at the single-neuron level may distinguish among the various computational algorithms which depend on error signals for their implementation.

Comparison to previous neurophysiology studies in the ventral stream

Prior work in IT has shown that responses of face-selective cells are stronger for atypical faces than for typical faces and are stronger when the parts are presented individually than when presented together in a whole face (35)(28). One possible interpretation of these data is that IT neurons are not tuned for faces but are instead tuned for atypical face features (i.e. extreme feature tuning) (35); however, our data suggest an alternative interpretation of this finding and, more deeply, an extended computational purpose of IT dynamics. In that prior work, the response preference of each neuron was determined by averaging over a long time window (~200 ms). By looking more closely at the fine time scale dynamics of the IT response, we suggest that this same “extreme coding” phenomenon can instead be interpreted as a natural consequence of networks that have an actual tuning preference for typical faces (as evidenced by an initial response preference for typical faces in pIT, cIT, and aIT; **Fig. 4b**) but that also compute error signals with respect to that preference. The hierarchical error coding framework proposed here provides a single, unifying account of many other reliable but previously unexplained phenomena in IT: sublinear integration of multiple inputs (35)(28) (**Fig. 6a**), tuning for extreme features in faces (35) (**Fig. 2**; positional shifts of the eye position correspond to extremes of the face space used in Freiwald et al., 2009), tuning for novel stimuli over familiar stimuli (36)(38)(37) (**Fig. 6b**, response to first presentation is larger for the novel inputs), and rapid response dynamics for familiar over novel images (13) (**Fig. 6b**, larger oscillation amplitude to repeated presentation of familiar inputs). Thus, we have provided a parsimonious framework that can account for these disparate neural phenomena by naturally extending the purely feedforward model in a way suggested by prior computational work. The perspective that many IT neurons code error signals reflecting deviations from the naturally learned statistics of images suggests that natural joint statistics of features will lead to suppressed responses on average in the hidden processing stages. This perspective may extend across the ventral visual stream including V1 where there is causal evidence of a suppressive role for feedback in producing end-stopping (44), and it has been suggested that end-stopping is the result of error-like computations (21).

Computational utility of coding errors in addition to states

In error-computing networks, errors provide control signals for guiding learning giving these networks additional adaptive power over basic feature estimation networks. This property helps augment the classical, feature coding view of neurons which, with only feature activations and Hebbian operations, does not lead to efficient gradient descent and may provide insight into how more intelligent unsupervised and supervised learning algorithms such as backpropagation could be plausibly implemented in the brain. A potentially important contribution of this work is the suggestion that gradient descent algorithms are facilitated by using an error code so that efficient learning is reduced to a simple Hebbian operation at synapses and efficient inference is simply integration of inputs at the cell body. This representational choice, to code the computational primitives of gradient descent in spiking activity, simply leverages existing neural machinery for inference and learning.

While we provide evidence that IT hidden units code error signals, the precise computational use of those error signals remains to be empirically determined. The most likely downstream, causal impact of error signals could be in online inference (updating neural firing rates), offline inference (updating synaptic weights, a.k.a. learning), or both. We expect that optimizing the weights and optimizing the states (feature estimates) are both implemented in cortical circuits as there is a large body of evidence on inference and learning in cortex including work in IT showing unsupervised learning of novel images within a few hundred presentations (37)(38)(45), work in IT showing neural changes after visual discrimination training (46), and work in IT showing that neural responses evolve over short timescales to produce improved estimates online (10)(11)(47). If both online and offline optimization mechanisms are used, their impact on the animal's behavioral output could ground their relative importance. Conversely, the animal's engagement in a task may guide the computation of error signals, through addition of a top-down supervision term as in classical error backpropagation, for improved performance under a new behavioral goal. As new tools become available to specifically disrupt cortico-cortical feedback, they can be brought to bear on these key questions in hierarchical cortical computation.

MATERIALS & METHODS

Animals and surgery. All surgery, behavioral training, imaging, and neurophysiological techniques are identical to those described in detail in our previous work[28]. Two rhesus macaque monkeys (*Macaca mulatta*) weighing 6 kg (Monkey 1, female) and 7 kg (Monkey 2, male) were used. A surgery using sterile technique was performed to implant a plastic fMRI compatible headpost prior to behavioral training and scanning. Following scanning, a second surgery was performed to implant a plastic chamber positioned to allow targeting of physiological recordings to posterior, middle, and anterior face patches in both animals. All procedures were performed in compliance with National Institutes of Health guidelines and the standards of the MIT Committee on Animal Care and the American Physiological Society.

Behavioral training and image presentation. Subjects were trained to passively fixate a central white fixation dot during serial visual presentation of images at a natural saccade-driven rate (one image every 200 ms). Although a 4° fixation window was enforced, subjects generally fixated a much smaller region of the image ($<1^o$)[28]. Images were presented at a size of 6° except for control tests at 3° and 12° sizes (**Fig. 3c**), and all images were presented for 100 ms duration with 100 ms gap (background gray screen) between each image. Up to 15 images were presented during a single fixation trial, and the first image presentation in each trial was discarded from later analyses. Five repetitions of each image in the general screen set were presented, and ten repetitions of each image were collected for all other image sets. The screen set consisted of a total of 40 images drawn from four categories (faces, bodies, objects, and places; 10 exemplars each) and was used to derive a measure of face versus nonface object selectivity. Following the screen set testing, some sites were tested using an image set containing images of face parts presented in different combinations and positions. We first segmented the face parts (eye, nose, mouth) from a monkey face image. These parts were then blended using a Gaussian window, and the face outline was filled with pink noise to create a continuous background texture. A face part could appear on the outline at any one of nine positions on an evenly spaced 3x3 grid. Although the number of possible images is large (4^9 = 262,144 images), we chose a subset of these images for testing neural sites (n=82 images).

588 Specifically, we tested the following images: the original whole face image, the noise-filled
589 outline, the whole face reconstructed by blending the four face parts with the outline, all possible
590 single part images where the eye, nose, or mouth could be at one of nine positions on the
591 outline (n=3x9=27 images), all two part images containing a nose, mouth, left eye, or right eye
592 at the correct outline-centered position and an eye tested at all remaining positions (n=4*8-
593 1=31images), all two part images containing a correctly positioned contralateral eye while
594 placing the nose or mouth at all other positions (n=2*8-2=14 images), and all correctly
595 configured faces but with one or two parts missing (n=3+4=7 images). The particular two-part
596 combinations tested were motivated by prior work demonstrating the importance of the eye in
597 early face processing[28], and we sought to determine how the position of the eye relative to the
598 outline and other face parts was encoded in neural responses. The three and four part
599 combinations were designed to manipulate the presence or absence of a face part for testing
600 the integration of face parts, and in these images, we did not vary the positions of the parts from
601 those in a naturally occurring face. In a follow-up test on a subset of sites, we permuted the
602 position of the four face parts under the constraint that they still formed the configuration of a
603 naturally occurring face (i.e. preserve the 'T' configuration, n=10 images; **Fig. 3b**). We tested
604 single part images at 3° and 12° sizes in a subset of sites (n=27 images at each size; **Fig. 3c**).
605 Finally, we measured the responses to the individual face parts in the absence of the outline
606 (n=4 images; **Fig. 6a**).
607
608 **MR Imaging and neurophysiological recordings.** Both structural and functional MRI scans
609 were collected in each monkey. Putative face patches were identified in fMRI maps of face
610 versus object selectivity in each subject. A stereo microfocal x-ray system[24] was used to guide
611 electrode penetrations in and around the fMRI defined face-selective subregions of IT. X-ray
612 based electrode localization was critical for making laminar assignments since electrode
613 penetrations are often not perpendicular to the cortical lamina when taking a doral-ventral
614 approach to IT face patches. Laminar assignments of recordings were made by co-registering x-
615 ray determined electrode coordinates to MRI where the pial-to-gray matter border and the gray-
616 to-white matter border were defined; based on our prior work estimating sources of error (e.g.
617 error from electrode tip localization and brain movement), registration of electrode tip locations
618 to MRI brain volumes has a total of <400 micron error which is sufficient to distinguish deep
619 from superficial layers[48]. Multi-unit activity (MUA) was systematically recorded at 300 micron
620 intervals starting from penetration of the superior temporal sulcus such that all sites were tested
621 with a screen set containing both faces and nonface objects, and a subset of sites that were
622 visually driven were further tested with our main image set manipulating the position of face
623 parts. Although we did not record single-unit activity, our previous work showed similar
624 responses between single-units and multi-units on images of the type presented here[28], and our
625 results are consistent with observations in previous single-unit work in IT[35]. Recordings were
626 made from PL, ML, and AM in the left hemisphere of monkeys 1 and 2 and from AL in monkey
627 2. AM and AL are pooled together in our analyses forming the aIT sample.
628
629 **Neural data analysis.** The face patches were physiologically defined in the same manner as in
630 our previous study[28]. Briefly, we fit a graded 3D sphere model (linear profile of selectivity that
631 rises from a baseline value toward the maximum at the center of the sphere) to the spatial
632 profile of face versus nonface object selectivity across our sites. We tested spherical regions
633 with radii from 1.5 to 10 mm and center positions within a 5 mm radius of the fMRI-based
634 centers of the face patches. The resulting physiologically defined regions were 1.5 to 3 mm in
635 diameter. Sites which passed a visual response screen (mean response in a 60-160 ms window
636 >2*SEM above baseline for at least one of the four categories in the screen set) were included

in further analysis. All firing rates were baseline subtracted using the activity in a 25-50 ms window following image onset averaged across all repetitions of an image. Finally, given that the visual response latencies in monkey 2 were on average 13 ms slower than those in monkey 1, we applied a single latency correction (13 ms shift to align monkey 1 and monkey 2's data) prior to averaging across monkeys. This was done to so as not to wash out any fine timescale dynamics by averaging though similar results were obtained without using this latency correction, and this single absolute adjustment was more straightforward than the site-by-site adjustment used in our previous work (similar results were obtained using this alternative latency correction)[28]. The observed selectivity dynamics (**Fig. 2**) were found in each monkey analyzed separately (**Fig. 3a**). Images that produced an average population response $\geq$ 0.9 of the initial response (60-100 ms) to a face-like image were analyzed further (**Figs. 2-4**). In follow-up analyses, we specifically limited comparison to images with the same number of parts (**Fig. 3b**). For example, for single part images, we used the image with the eye in the upper, contralateral region of the outline as a reference and found that four other images of the 27 single-part images elicited a response at least as large as 90% of the response to this standard image. For images containing all four face parts, we used the whole face as the standard and found nonface-like arrangements of the four face parts that drove at least 90% of the early response to the whole face (2 images out of 10 tested). Individual site d' measures were computed using $d' = (u_1 - u_2)/(var_1 + var_2)/2)^{1/2}$ where variance was computed across all trials for that image class (i.e. all presentations of all typical face images). A positive d' implies a stronger response to more naturally occurring (typical) arrangements of face parts while a negative d' indicates a preference for unnatural (atypical) arrangements of the face parts.

Dynamical models

Modeling framework and equations. To model the dynamics of neural response rates in a hierarchy, we started with the simplest possible model that might capture those dynamics: we used a model architecture consisting of a hidden stage of processing containing two units that linearly converged onto a single output unit. An external input was applied separately to each hidden stage unit, which can be viewed as representing different features for downstream integration. We varied the connections between the two hidden units within the hidden processing stage (lateral connections) or between hidden and output stage units (feedforward and feedback connections) to instantiate different model families. The details of the different architectures specified by each model class can be visualized by their equivalent neural network diagrams and connection matrices (**Supplementary Fig. 1**). Here, we provide a basic description for each model tested. All models utilize a 2x2 feedforward identity matrix *A* that simply transfers inputs ***u*** (2x1) to hidden layer units ***x*** (2x1) and a 1x2 feedforward matrix *B* that integrates hidden layer activations ***x*** into a single output unit *y*.

$$A = k_a \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad B = k_b [1 \quad 1] \tag{1}$$

To generate dynamics in the simple networks below, we assumed that neurons act as leaky integrators of their total synaptic input, a standard rate-based model of a neuron used in previous work[14,21].

Pure feedforward. In the purely feedforward family, connections were exclusively from hidden to output stages through feedforward matrices *A* and *B*.

$$\dot{\mathbf{x}} = A\mathbf{u} - \mathbf{x}/\tau \qquad \dot{y} = B\mathbf{x} - y/\tau \tag{2}$$

where τ is the time constant of the leak current which can be seen as reflecting the biophysical limitations of neurons (a perfect integrator with large τ would have almost no leak and hence infinite memory).

Lateral inhibition. Lateral connections (matrix with off-diagonal terms) are included and are inhibitory. The scalar k_l sets the relative strength of lateral inhibition versus bottom-up input.

$$\dot{\mathbf{x}} = A\mathbf{u} - \begin{bmatrix} 0 & k_l \\ k_l & 0 \end{bmatrix}\mathbf{x} - \mathbf{x}/\tau \qquad \dot{y} = B\mathbf{x} - y/\tau \tag{3}$$

Normalization. An inhibitory term that scales with the summed activity of units within a stage is included. The scalar k_s sets the relative strength of normalization versus bottom-up input.

$$\dot{\mathbf{x}} = A\mathbf{u} - k_s \sum x \cdot \mathbf{x} - \mathbf{x}/\tau \qquad \dot{y} = B\mathbf{x} - k_s y \cdot y - y/\tau \tag{4}$$

Normalization (nonlinear)[23]*.* The summed activity of units within a stage is used to nonlinearly scale shunting inhibition.

$$\dot{\mathbf{x}} = A\mathbf{u} - \frac{\mathbf{x}}{\tau\sqrt{1 - k_s \sum x}} \qquad \dot{y} = B\mathbf{x} - \frac{y}{\tau\sqrt{1 - k_s y}} \tag{5}$$

Since the normalization term in equation (5) is not continuously differentiable, we used the fourth-order Taylor approximation around zero in the simulations of equation (5).

Feedback (linear reconstruction). The feedback-based model is derived using a normative framework that performs optimal inference in the linear case[14] (unlike the networks in equations (2)-(5) which are motivated from a mechanistic perspective but do not directly optimize a squared error performance loss). The feedback network minimizes the cost *C* of reconstructing the inputs of each stage (i.e. mean squared error of layer *n* predicting layer *n-1*).

$$C = \frac{1}{2}\left(\mathbf{u} - A^T\mathbf{x}\right)^2 + \frac{1}{2}\left(\mathbf{x} - B^T y\right)^2 \tag{6}$$

Differentiating this coding cost with respect to the encoding variables in each layer ***x***, *y* yields:

$$\frac{\partial C}{\partial \mathbf{x}} = -A(\mathbf{u} - A^T\mathbf{x}) + (\mathbf{x} - B^T\mathbf{y}) \qquad \frac{\partial C}{\partial \mathbf{y}} = -B(\mathbf{x} - B^T\mathbf{y}) \tag{7}$$

The cost function *C* can be minimized by descending these gradients over time to optimize the values of ***x*** and *y*:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = -\frac{\partial C}{\partial \mathbf{x}} = A(\mathbf{u} - A^T\mathbf{x}) - (\mathbf{x} - B^T\mathbf{y}) - \mathbf{x}/\tau \tag{8}$$

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} = -\frac{\partial C}{\partial \mathbf{y}} = B(\mathbf{x} - B^T\mathbf{y}) - \mathbf{y}/\tau$$

The above dynamical equations are equivalent to a linear network with a connection matrix containing symmetric feedforward (B) and feedback (B^T) weights between stages $\boldsymbol{x}$ and y as well as within-stage pooling followed by recurrent inhibition ($-AA^T\boldsymbol{x}$ and $-BB^Ty$) that resembles normalization. The property that symmetric connections minimize the cost function C generalizes to a feedforward network of any size or number of hidden processing stages (i.e. holds for arbitrary lower triangular connection matrix B). The final activation states ($\boldsymbol{x}$,y) of the hierarchical generative network are optimal in the sense that the bottom-up activations (implemented through feedforward connections) are balanced by the top-down expectations (implemented by feedback connections) which is equivalent to a Bayesian network combining bottom-up likelihoods with top-down priors to compute the maximum *a posteriori* (MAP) estimate. Here, the priors are embedded in the weight structure of the network. In simulations, we include an additional scalar k_{td} that sets the relative weighting of bottom-up versus top-down signals.

$$\dot{\mathbf{x}} = A(\mathbf{u} - A^T\mathbf{x}) - k_{td}(\mathbf{x} - B^Ty) - \mathbf{x}/\tau \qquad (9)$$

Error signals computed in the feedback model. In equation (9), inference can be thought of as proceeding through integration of inputs on the dendrites of neuron population $\boldsymbol{x}$. In this scenario, all computations are implicit in dendritic integration. Alternatively, the computations in equation (9) can be done in two steps where, in the first step, reconstruction errors are computed (i.e. $\boldsymbol{e_0}=\boldsymbol{u}-A^T\boldsymbol{x}$, $\boldsymbol{e_1}=\boldsymbol{x}-B^Ty$) and explicitly represented in a separate error coding population. These error signals can then be integrated to generate the requisite update to the state signal of neuron population $\boldsymbol{x}$.

$$\dot{\mathbf{x}} = A\mathbf{e}_0 - k_{td}\mathbf{e}_1 - \mathbf{x}/\tau \qquad \dot{\mathbf{y}} = B\mathbf{e}_1 - y/\tau \qquad (10)$$

An advantage of this strategy is that there are now only two input populations to a state unit, and those inputs allow implementation of an efficient Hebbian rule for learning weight matrices[21] -- the gradient rule for learning is simply a product of the state activation and the input error activation (weight updates obtained by differentiating equation (6) with respect to weight matrices A and B: $\Delta A = \mathbf{x}\cdot\boldsymbol{e_0}^T$, $\Delta A^T = \boldsymbol{e_0}\cdot\mathbf{x}^T$, $\Delta B = y\cdot\boldsymbol{e_1}^T$, and $\Delta B^T = \boldsymbol{e_1}\cdot y$). Thus, the reconstruction errors serve as computational intermediates for both the gradients of online inference (dynamics in state space, equation (10)) and gradients for offline learning (dynamics in weight space).

In order for the reconstruction errors at each layer to be scaled appropriately in the feedback model, we invoke an additional downstream variable z to predict activity at the top layer such that, instead of $\boldsymbol{e_2}=y$ which scales as a state variable, we have $\boldsymbol{e_2}=y-C^Tz$ (**Supplementary Fig. 1a**). This overall model reflects a state and error coding model as opposed to a state only model. A third, less plausible possibility is to only represent error signals explicitly (a.k.a. pure predictive coding; **Fig. 7** (iii)). In other words, the linear dynamical system in equation (8) can be rewritten through a linear change of variables to include error variables only, and the state variables become implicitly represented during dendritic integration of these errors.

Feedback (three-stage). For the simulations in **Figs. 5,6**, a three-stage version of the above

models was used. These deeper network were also wider such that they began with four input units (***u***) instead of only two inputs in the two-stage models. These inputs converged through successive processing stages (***w,x,y***) to one unit at the top node (***z***) (**Supplementary Fig. 1b**).

Feedback (nonlinear reconstruction). We tested versions of feedback-based models that optimized different cost functions other than a linear reconstruction cost (**Supplementary Fig. 2**). In nonlinear hierarchical inference, reconstruction is performed using a monotonic nonlinearity with a threshold (*th*) and bias (*bi*):

$$C = \frac{1}{2}\left(\mathbf{u} - f(A^T\mathbf{x})\right)^2 + \frac{1}{2}\left(\mathbf{x} - f(B^T\mathbf{y})\right)^2, \quad where\ f(x) = tanh(x - th) + bi \tag{11}$$

$$\begin{aligned} \dot{\mathbf{x}} &= A(\mathbf{u} - f(A^T\mathbf{x}))(1 - tanh(A^T\mathbf{x} - th)^2) - k_{td}(\mathbf{x} - f(B^T y)) - \mathbf{x}/\tau \\ \dot{\mathbf{y}} &= B(\mathbf{x} - f(B^T\mathbf{y}))(1 - tanh(B^T\mathbf{y} - th)^2) - \mathbf{y}/\tau \end{aligned} \tag{12}$$

Feedback (linear construction). Instead of a reconstruction cost, we additionally simulated the states and errors in a feedback network minimizing a linear construction cost:

$$C = \frac{1}{2}\left(A\mathbf{u} - \mathbf{x}\right)^2 + \frac{1}{2}\left(B\mathbf{x} - \mathbf{y}\right)^2 \tag{13}$$

$$\dot{\mathbf{x}} = (A\mathbf{u} - \mathbf{x}) - k_{td}B^T(B\mathbf{x} - \mathbf{y}) - \mathbf{x}/\tau \qquad \dot{\mathbf{y}} = (B\mathbf{x} - \mathbf{y}) - \mathbf{y}/\tau \tag{14}$$

Model simulation. To simulate the dynamical systems in equations (2)-(14), a step input ***u*** was applied. This input was smoothed using a Gaussian kernel to approximate the lowpass nature of signal propagation in the series of processing stages from the retina to pIT:

$$\mathbf{u}(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(t-t_0)^2}{2\sigma^2}} * \mathbf{h}(t) \Rightarrow \dot{\mathbf{u}} = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(t-t_0)^2}{2\sigma^2}} \cdot \mathbf{h} \tag{15}$$

where the elements of ***h*** are scaled Heaviside step functions. The input is thus a sigmoidal ramp whose latency to half height is set by t_0 and rise time is set by σ. For simulation of two-stage models, there were ten basic parameters: latency of the input t_0, standard deviation of the Gaussian ramp σ, system time constant τ, input connection strength A, feedforward connection strength B, the four input values across two stimulus conditions (i.e. h_{11}, h_{12}, h_{21}, h_{22}), and a factor *sc* for scaling the final output to the neural activity. In the deeper three-stage network, there were a total of fifteen parameters which included an additional feedforward connection strength C and additional input values since the three-stage model had four inputs instead of two. The lateral inhibition model class required one additional parameter k_l as did the normalization model family k_s, and for feedback model simulations, there was an additional feedback weight k_{td} to scale the relative contribution of the top-down errors in driving online inference. For the error coding variants of the feedback model, gain parameters C (two-stage) and D (three-stage) were included to scale the overall magnitude of the top level reconstruction error.

Model parameter fits to neural data. In fitting the models to the observed neural dynamics, we mapped the summed activity in the hidden stage (***x***) to population averaged activity in pIT, and we mapped the summed activity in the output stage (***y***) to population averaged signals measured in aIT. To simulate error coding, we mapped the reconstruction errors $\boldsymbol{e_1}=\boldsymbol{x}-B^T y$ and $\boldsymbol{e_2}=y-C^T z$ to activity in pIT and aIT, respectively. We applied a squaring nonlinearity to the model outputs as an approximation to rectification since recorded extracellular firing rates are non-negative (and linear rectification is not continuously differentiable). Analytically solving this system of dynamical equations (2)-(14) for a step input is precluded because of the higher order interaction terms (the roots of the determinant and hence the eigenvalues/eigenvectors of a 3x3 matrix are not analytically determined, except for the purely feedforward model which only has first-order interactions), and in the case of the normalization models, there is an additional nonlinear dependence on the shunt term. Thus, we relied on computational methods (constrained nonlinear optimization) to fit the parameters of the dynamical systems to the neural data with a quadratic (sum of squares) loss function.

Parameter values were fit in a two step procedure. In the first step, we fit only the difference in response between image classes (differential mode which is the selectivity profile over time, see **Fig. 5b**, left panel), and in the second step, we refined fits to capture an equally weighted average of the differential mode and the common mode (the common mode is the average across images of the response time course of visual drive). This two-step procedure was used to ensure that each model had the best chance of fitting the dynamics of selectivity (differential mode) as these selectivity profiles were the main phenomena of interest but were smaller in size (20% of response) compared to overall visual drive. In each step, fits were done using a large-scale algorithm (interior-point) to optimize coarsely, and the resulting solution was used as the initial condition for a medium-scale algorithm (sequential quadratic programming) for additional refinement. The lower and upper parameter bounds tested were: t_0=[50 70], σ=[0.5 25], τ =[0.5 1000], k_l,k_s,k_{td}=[0 1], A,B,C,D=[0 2], h=[0 20], sc=[0 100], th=[-20 20], and bi=[-1 1] which proved to be adequately liberal as parameter values converged to values that did not generally approach these boundaries. To avoid local minima, the algorithm was initialized to a number of randomly selected points (n=50), and after fitting the differential mode, we took the top fits (n=25) for each model class and used these as initializations in subsequent steps. The single best fitting instance of each model class is shown in the main figures.

Model predictions. For the predictions in **Fig. 6**, all architectural parameters obtained by the fitting procedure above were held fixed; only the pattern of inputs to the network was varied. For **Fig. 6a**, to test the input integration properties of a model, we used the top-performing model (most optimal solution) and compared the response to all inputs presented simultaneously with the sum of the responses to each input alone.

For **Fig. 6b**, we approximated novel versus familiar images as random patterns versus structured input patterns that matched the learned weights of the network. Here, we used a version of the model with two independent outputs reflecting detectors for two familiarized input patterns (output 1 tuned to pattern A: u_1, u_2, u_3, u_4 active and output 2 tuned to pattern 2: u_5, u_6, u_7, u_8 active) (**Fig. 6b**). Alternating between these two input patterns simulates alternation of two familiarized (learned) images as compared to purely random patterns ($u_{1\text{-}8}$ independent and identically distributed). To parametrically vary the degree of correlation of inputs to the learned weight patterns from random (correlation = 0) to deterministic (correlation = 1), we drew input values from a joint distribution $P(u_1,u_2,u_3,u_4,u_5,u_6,u_7,u_8)$ where $u_{1\text{-}4}$ were drawn from a high-

valued uniform distribution on the interval [1-ε,1] and $u_{5\text{-}8}$ were drawn from a low-valued uniform distribution [0, ε] for stimulus pattern *A* and the opposite for pattern *B* ($u_{5\text{-}8}$ high-valued and $u_{1\text{-}4}$ low-valued). The parameter ε determines the range of values that could be drawn from purely deterministic (0 or 1) to randomly uniformly distributed (from 0 to 1). Thus, the correlation of the inputs correspondingly varies according to $\rho(u_i,u_j)= \rho(u_k,u_l)=(\varepsilon\text{-}1)^2/((\varepsilon\text{-}1)^2+\varepsilon^2/3)$ where $1\leq i,j\leq 4$, $i\neq j$ and $5\leq k,l\leq 8$, $k\neq l$ approaching correlation equal to 0 for a purely, random pattern (ε=1) that had a low probability of matching the learned patterns *A* and *B*.

Code availability. All data analysis and computational modeling were done using custom scripts written in Matlab. All code is available upon request.

Statistics. Error bars represent standard errors of the mean obtained by bootstrap resampling (n = 1000). All statistical comparisons including those of means or correlation values were based on 99% confidence intervals obtained by bootstrap resampling (n = 1000). All statistical tests were two-sided unless otherwise specified. Spearman's rank correlation coefficient was used.

ACKNOWLEDGEMENTS

We thank J. Deutsch, K. Schmidt, and P. Aparicio for help with MRI and animal care and B. Andken and C. Stawarz for help with experiment software. We are grateful to T. Meyer and C. Olson for sharing figures from their published work. This research was supported by US National Eye Institute grants R01-EY014970 (J.J.D.) and K99-EY022671 (E.B.I.), NRSA postdoctoral fellowships F32-EY019609 (E.B.I.) and F32-EY022845-01 (C.F.C.), Office of Naval Research MURI-114407 (J.J.D), and The McGovern Institute for Brain Research.

REFERENCES

1. Hung CP, Kreiman G, Poggio T, DiCarlo JJ (2005) Fast Readout of Object Identity from Macaque Inferior Temporal Cortex. *Science* 310(5749):863–866.

2. Majaj NJ, Hong H, Solomon EA, DiCarlo JJ (2015) Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *J Neurosci* 35(39):13402–13418.

3. Pasupathy A, Connor CE (2001) Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation. *J Neurophysiol* 86(5):2505–2519.

4. Rust NC, DiCarlo JJ (2010) Selectivity and Tolerance ("Invariance") Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *J Neurosci* 30(39):12978–12995.

5. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, pp 1106–1114.

6. Zeiler MD, Fergus R (2013) Stochastic Pooling for Regularization of Deep Convolutional Neural Networks. *ArXiv13013557 Cs Stat*. Available at: http://arxiv.org/abs/1301.3557 [Accessed February 12, 2016].

7. Yamins DLK, et al. (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc Natl Acad Sci* 111(23):8619–8624.

8. Cadieu CF, et al. (2014) Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLoS Comput Biol* 10(12):e1003963.

9. DiCarlo JJ, Zoccolan D, Rust NC (2012) How Does the Brain Solve Visual Object Recognition? *Neuron* 73(3):415–434.

10. Brincat SL, Connor CE (2006) Dynamic Shape Synthesis in Posterior Inferotemporal Cortex. *Neuron* 49(1):17–24.

11. Sugase Y, Yamane S, Ueno S, Kawano K (1999) Global and fine information coded by single neurons in the temporal visual cortex. *Nature* 400(6747):869–873.

12. Chen M, et al. (2014) Incremental Integration of Global Contours through Interplay between Visual Cortical Areas. *Neuron* 82(3):682–694.

13. Meyer T, Walker C, Cho RY, Olson CR (2014) Image familiarization sharpens response dynamics of neurons in inferotemporal cortex. *Nat Neurosci* 17(10):1388–1394.

14. Seung HS (1997) Pattern analysis and synthesis in attractor neural networks. *Theor Asp Neural Comput Multidiscip Perspect Singap*.

15. Fiser J, Berkes P, Orbán G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn Sci* 14(3):119–130.

16. Lee TS, Yang CF, Romero RD, Mumford D (2002) Neural activity in early visual cortex reflects behavioral experience and higher-order perceptual saliency. *Nat Neurosci* 5(6):589–597.

17. Zhang NR, Heydt R von der (2010) Analysis of the Context Integration Mechanisms Underlying Figure–Ground Organization in the Visual Cortex. *J Neurosci* 30(19):6482–6496.

18. Epshtein B, Lifshitz I, Ullman S (2008) Image interpretation by a single bottom-up top-down cycle. *Proc Natl Acad Sci* 105(38):14298–14303.

19. Williams DRGHR, Hinton GE (1986) Learning representations by back-propagating errors. *Nature* 323:533–536.

20. Olshausen BA, Field DJ (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609.

21. Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2(1):79–87.

22. Cavanaugh JR, Bair W, Movshon JA (2002) Nature and Interaction of Signals From the Receptive Field Center and Surround in Macaque V1 Neurons. *J Neurophysiol* 88(5):2530–2546.

23. Carandini M, Heeger DJ, Movshon JA (1997) Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. *J Neurosci* 17(21):8621–8644.

24. Cox DD, Papanastassiou AM, Oreper D, Andken BB, DiCarlo JJ (2008) High-Resolution Three-Dimensional Microelectrode Brain Mapping Using Stereo Microfocal X-ray Imaging. *J Neurophysiol* 100(5):2966–2976.

25. Tsao DY, Freiwald WA, Tootell RBH, Livingstone MS (2006) A Cortical Region Consisting Entirely of Face-Selective Cells. *Science* 311(5761):670–674.

26. Afraz S-R, Kiani R, Esteky H (2006) Microstimulation of inferotemporal cortex influences face categorization. *Nature* 442(7103):692–695.

27. Adelson EH, Movshon JA (1982) Phenomenal coherence of moving visual patterns. *Nature* 300(5892):523–525.

28. Issa EB, DiCarlo JJ (2012) Precedence of the Eye Region in Neural Processing of Faces. *J Neurosci* 32(47):16666–16682.

29. Freiwald WA, Tsao DY (2010) Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science* 330(6005):845–851.

30. Egeth HE, Yantis and S (1997) VISUAL ATTENTION: Control, Representation, and Time Course. *Annu Rev Psychol* 48(1):269–297.

31. Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A* 20(7):1434.

32. Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for boltzmann machines. *Cogn Sci* 9(1):147–169.

33. Rifai S, Vincent P, Muller X, Glorot X, Bengio Y (2011) Contractive auto-encoders: Explicit invariance during feature extraction. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp 833–840.

34. Meyers EM, Borzello M, Freiwald WA, Tsao D (2015) Intelligent Information Loss: The Coding of Facial Identity, Head Pose, and Non-Face Information in the Macaque Face Patch System. *J Neurosci* 35(18):7069–7081.

35. Freiwald WA, Tsao DY, Livingstone MS (2009) A face feature space in the macaque temporal lobe. *Nat Neurosci* 12(9):1187–1196.

36. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2006) Experience-Dependent Sharpening of Visual Shape Selectivity in Inferior Temporal Cortex. *Cereb Cortex* 16(11):1631–1644.

37. Woloszyn L, Sheinberg DL (2012) Effects of Long-Term Visual Experience on Responses of Distinct Classes of Single Units in Inferior Temporal Cortex. *Neuron* 74(1):193–205.

38. Meyer T, Olson CR (2011) Statistical learning of visual transitions in monkey inferotemporal cortex. *Proc Natl Acad Sci* 108(48):19401–19406.

39. Murray EA, Bussey TJ, Saksida LM (2007) Visual Perception and Memory: A New View of Medial Temporal Lobe Function in Primates and Rodents. *Annu Rev Neurosci* 30(1):99–122.

40. Friston K, Kiebel S (2009) Cortical circuits for perceptual inference. *Neural Netw* 22(8):1093–1104.

41. Hyvärinen A, Hurri J, Hoyer PO (2009) Lateral interactions and feedback. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision.* (Springer Science & Business Media), pp 307–318.

42. Salakhutdinov R, Hinton G (2012) An Efficient Learning Procedure for Deep Boltzmann Machines. *Neural Comput* 24(8):1967–2006.

43. Patel AB, Nguyen T, Baraniuk RG (2015) A Probabilistic Theory of Deep Learning. *ArXiv150400641 Cs Stat*. Available at: http://arxiv.org/abs/1504.00641 [Accessed February 14, 2016].

44. Nassi JJ, Lomber SG, Born RT (2013) Corticocortical Feedback Contributes to Surround Suppression in V1 of the Alert Primate. *J Neurosci* 33(19):8504–8517.

45. Li N, DiCarlo JJ (2008) Unsupervised Natural Experience Rapidly Alters Invariant Object Representation in Visual Cortex. *Science* 321(5895):1502–1507.

46. Baker CI, Behrmann M, Olson CR (2002) Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. *Nat Neurosci* 5(11):1210–1216.

47. Tang H, et al. (2014) Spatiotemporal Dynamics Underlying Object Completion in Human Ventral Visual Cortex. *Neuron* 83(3):736–748.

48. Issa EB, Papanastassiou AM, DiCarlo JJ (2013) Large-Scale, High-Resolution Neurophysiological Maps Underlying fMRI of Macaque Temporal Lobe. *J Neurosci* 33(38):15207–15219.

Figure 1

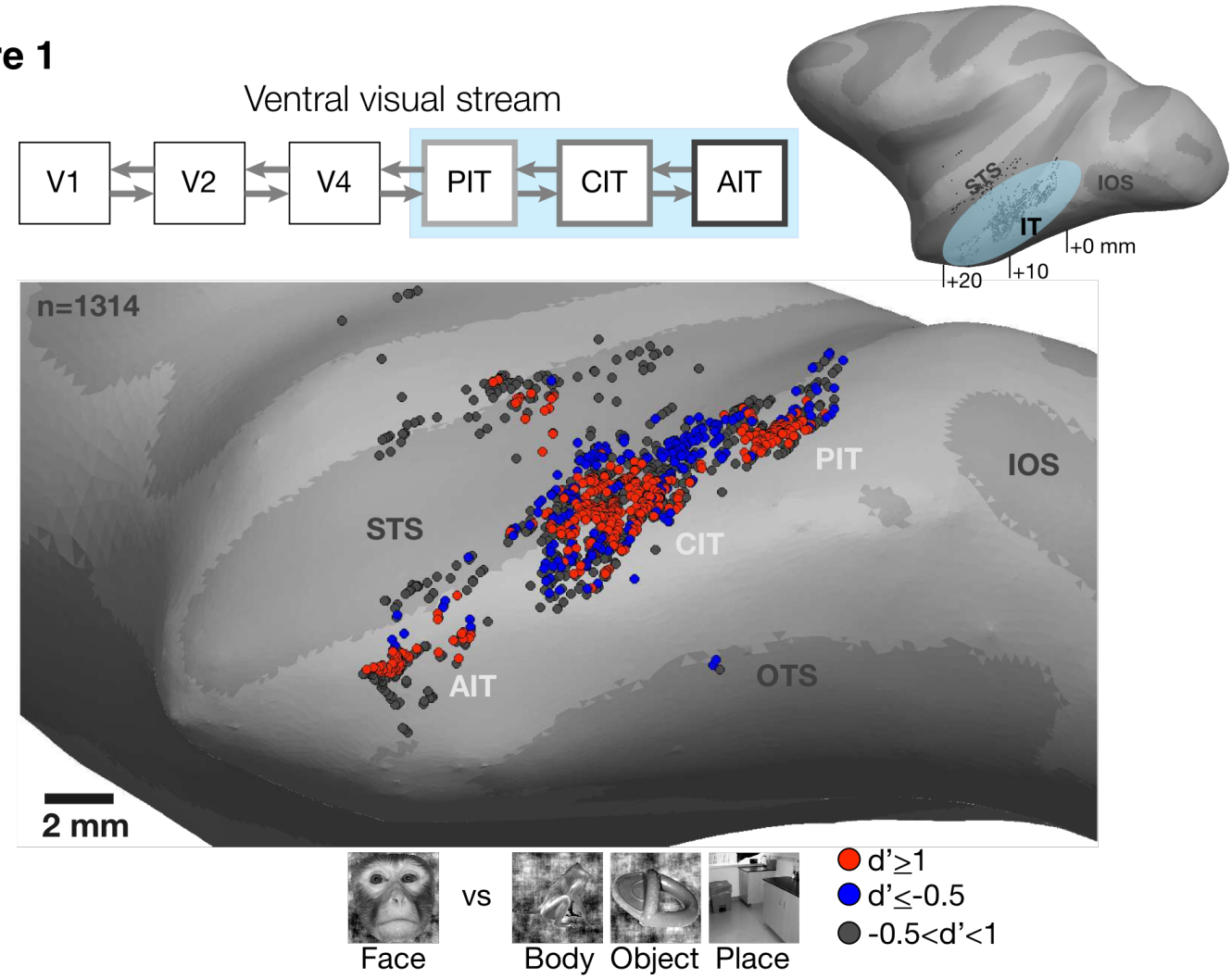


Figure 1 Neurophysiological recordings of face-selective subregions in the ventral visual stream. The ventral visual stream is a series of hierarchically connected areas (diagram in top row) and includes at least three IT processing stages (blue box). Neurons in these three stages were recorded along the lateral convexity of the inferior temporal lobe spanning the posterior to anterior extent of IT (+0 to +20 mm AP, Horsely-Clarke coordinates) in two monkeys (data from monkey 1 are shown). "Face neuron" sites (red) were operationally defined as those with a response preference for images of faces versus images of nonface objects (see **SI Methods**). While these were found throughout IT, they tended to be found in clusters that mapped to previously identified subdivisions of IT (posterior, central, and anterior IT) and also corresponded to face-selective areas identified under fMRI in the same subject (28)(48) (STS=superior temporal sulcus, IOS=inferior occipital sulcus, OTS=occipitotemporal sulcus).

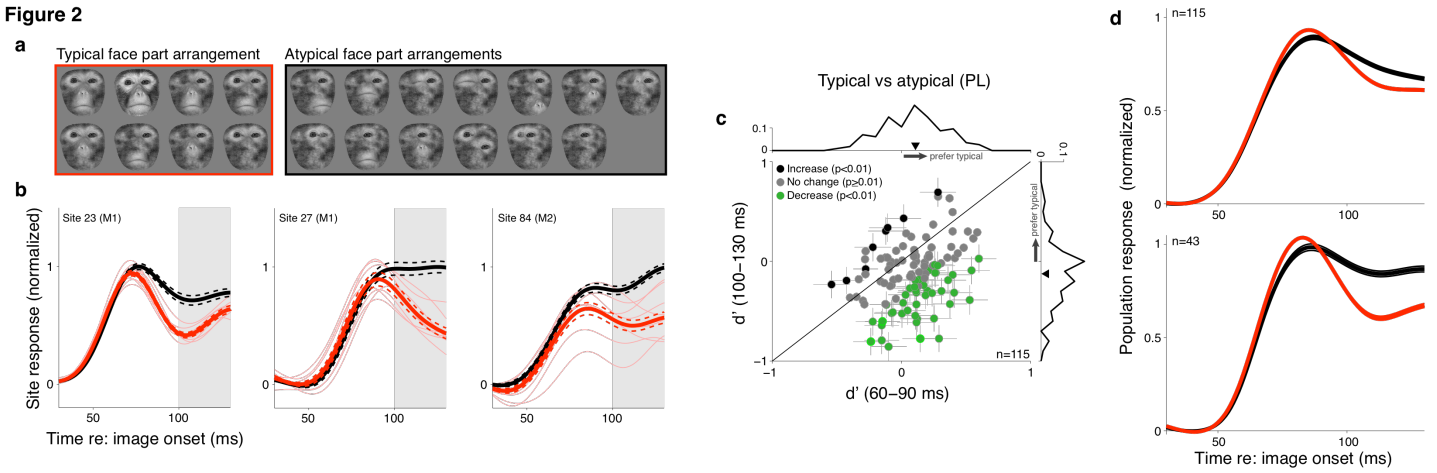


Figure 2 Responses in posterior IT to face-like images with typical and atypical face part arrangements. **(a)** Novel images were generated from an exemplar monkey face by positioning the face parts in different positions within the face outline. This procedure generated both typical (red) and atypical (black) arrangements of the face parts, and images that drove ≥90% of the early (60-100 ms) response to the whole face at the population level are shown here (first image in red box is synthesized whole face; compare to the second image which is the original whole face). **(b)** Responses of three example sites. Preference for images with atypical arrangements (black line = mean responses of 13 images shown in **(a)**) over images with typical arrangements (red lines; 8 images) emerged late in the neural response (gray box). **(c)** Preference for typical versus atypical arrangements of individual sites in early (60-90 ms) and late (100-130 ms, same as gray box in **(b)**) time windows (sites showing significant changes in their preference ($p<0.01$) are plotted with error bars and colored according to whether their preference (d') for typical vs. atypical face images increased (black), decreased (green), or did not change (gray); the three example sites from **(b)** all had decreasing preference and are highlighted by the green outlines; the marginal distributions are shown where inverted triangles are the median d' in early and late time windows). **(d)** Population average response in pIT across all images from **(a)** with typical or atypical arrangements for the whole recorded population (top) and for the sites showing significantly decreased preference for typical versus atypical arrangements (bottom) (same sites as green dots in **(c)**).

Figure 3

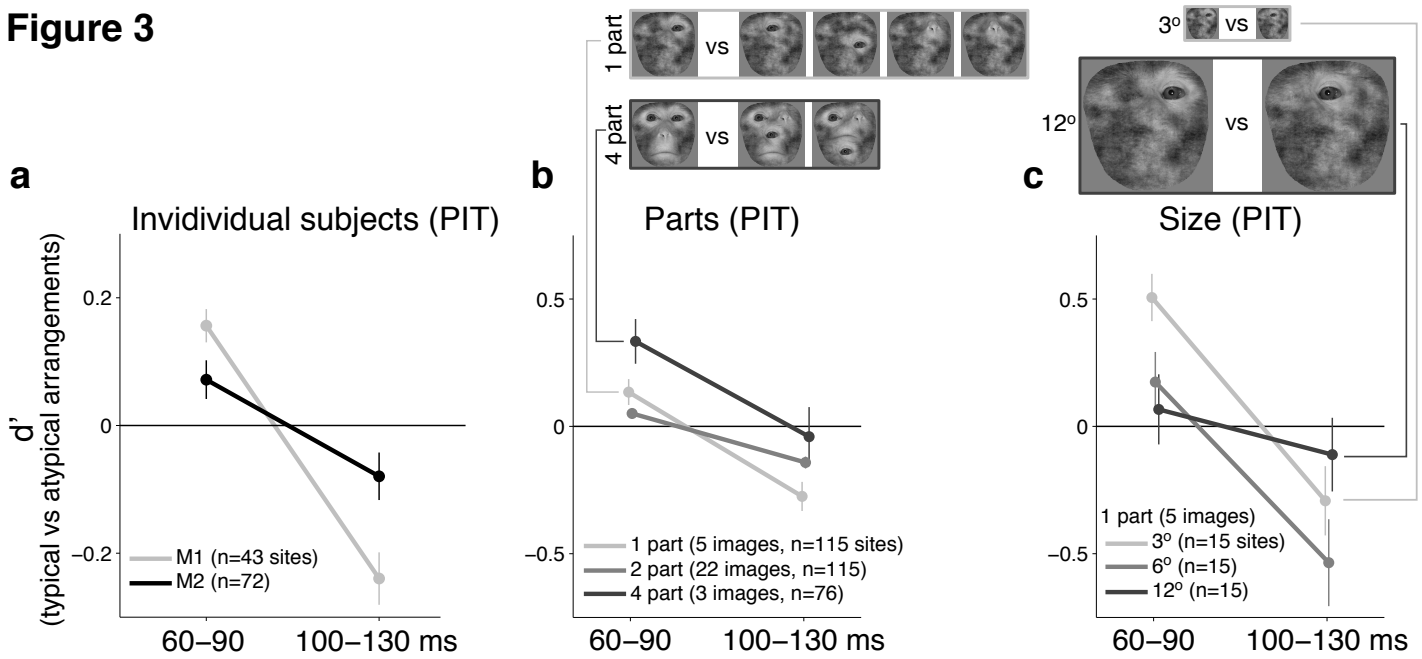


Figure 3 Individual monkey comparison and image controls for the reversal of rank-order selectivity in pIT. **(a)** Preference for images with typical versus atypical face part arrangements analyzed separately for each monkey. Median d' across sites in both early and late time windows is shown. **(b)** Preference for images with typical versus atypical face part arrangements was re-computed using image subsets containing the same number of parts in the outline (the five 1-part and the three 4-part image subsets shown at top; the larger 2-part subset contained 30 images total and is not shown). **(c)** The 1-part image subset was further tested at three different sizes (3°, 6°, and 12°). In all cases, pIT responses showed an initial preference for typically-arranged face parts, followed by a later preference for atypically arranged face parts.

Figure 4

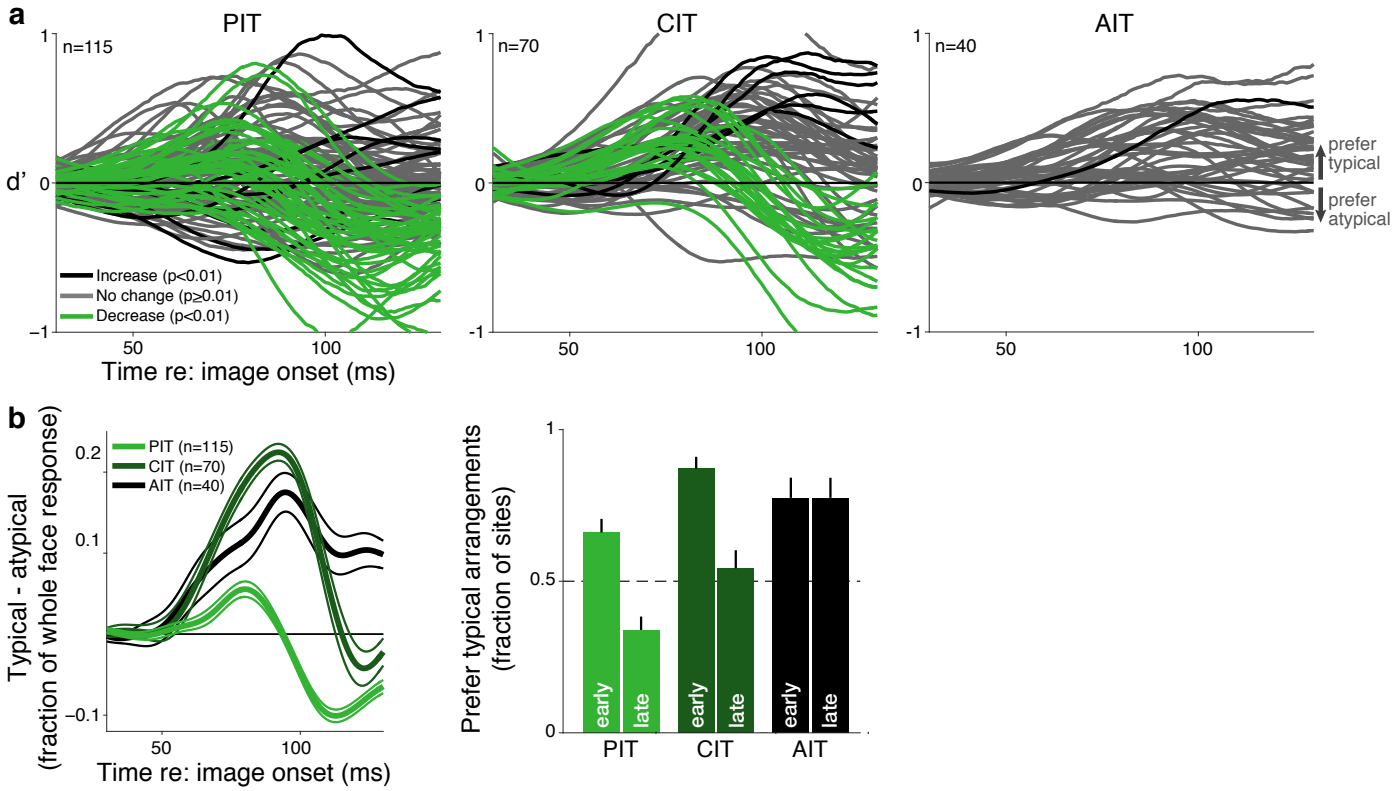


Figure 4 Time course of neural response preferences in posterior, central, and anterior IT for images with typical versus atypical face part arrangements. **(a)** The initial preference (d') in pIT for typical face part arrangements reversed in many sites following the first 30 ms of the response (left panel; sites in green showed significant decreases in selectivity ($p<0.01$); same color scheme as **Fig. 2c**). In aIT, selectivity for face-like arrangements of parts generally increased over time (right panel) (sites plotted in order of absolute change in preference; sites with largest absolute change in preference are in foreground). **(b)** Left: Population average response to typical versus atypical arrangements (left, normalized by the mean response of the population to the whole face). Right: The fraction of sites whose responses showed a preference for images of typical, face-like arrangements of the face parts (right) in pIT (green), cIT (dark green) and aIT (black) in the early (60-90 ms) and late (100-130 ms) phase of the response. Note that, in the late phase of the response, most pIT neurons paradoxically show a preference for atypical arrangements of face parts.

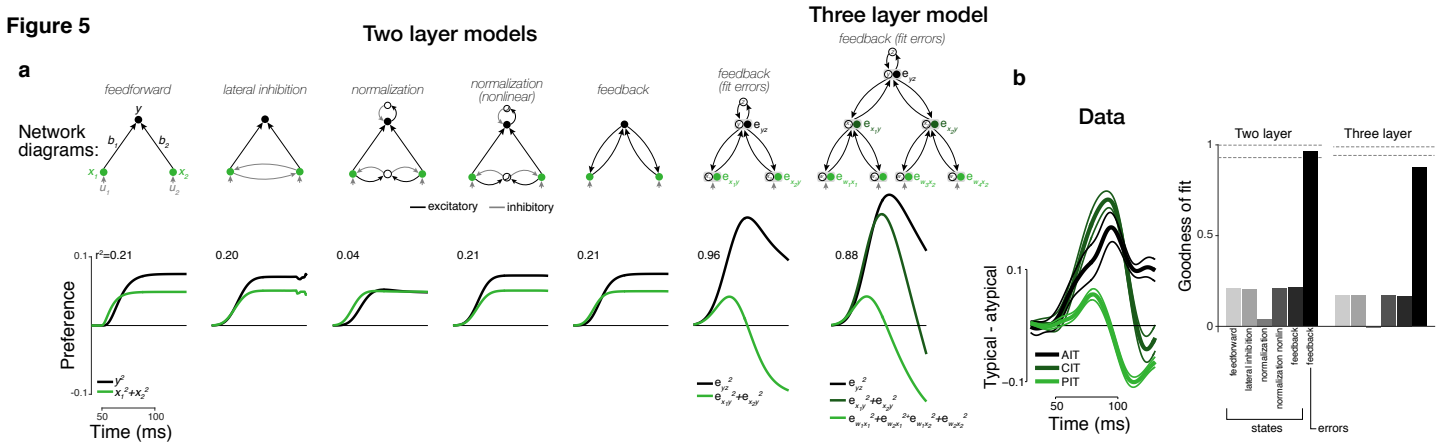


Figure 5 Computational modeling of neural dynamics in IT. **(a)** Simple two-layer neural networks (first six columns) with recurrent dynamics (network diagrams in top row; see **SI Methods** and **Supplementary Fig. 1 a,b**) were constructed to model neural signals measured in pIT and aIT corresponding to the first (green) and second (black) model processing stages. All models received two inputs (gray) into two hidden stage units (green) which sent feedforward projections that converged onto a single unit in the output stage (black). Besides this feedforward architecture, additional excitatory and inhibitory connections between units were used to implement recurrent dynamics (self connections reflecting leak currents are not shown here for clarity; see **Supplementary Fig. 1a** for detailed diagrams). In the five models on the left, the responses of the simulated neurons are assumed to code the current estimates of some set of features in the world (a.k.a states), as is standard in most such networks. The best fit to the population averaged neural data in **(b) (**same as left panel in **Fig. 4b**) of the states of each model class are shown (first five columns). These state coding models showed increasing selectivity over time from hidden to output layers and did not demonstrate the strong reversal of stimulus preference in their hidden processing stage (green lines) as observed in the pIT neural population. However, the neurons coding errors in a feedback-based hierarchical model did show a strong reversal of stimulus preference in the hidden processing stage (sixth column; reconstruction errors instead of the states were fit directly to the data). This model which codes the error signals (filled circles) also codes the states (open circles) (network diagram in sixth column of top row). A three-layer version of this model also produced reversal dynamics in the hidden layers (seventh column). **(b)** Population averaged neural selectivity profile shown (left, same as left panel in **Fig.** 4b). (right) Goodness of fit of all two-layer and three-layer models tested (dashed lines represent mean and standard error of reliability of neural data as estimated by bootstrap resampling).

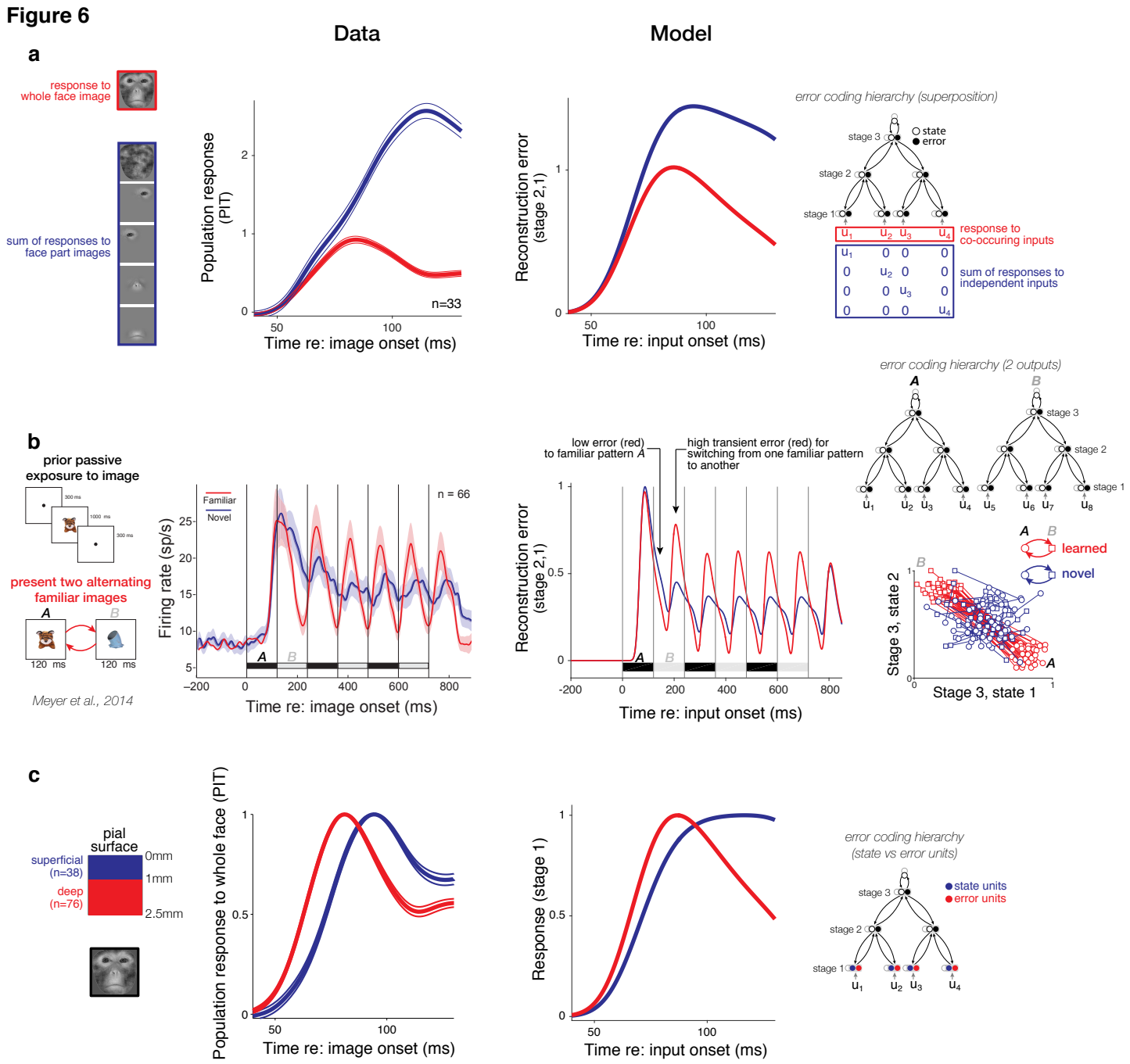


Figure 6 Comparison of neuronal findings and predictions of a feedback-based error coding model. To generate model predictions, the architectural parameters of the model (i.e. connection weights and time constants) were held fixed, and only the input patterns ***u*** were varied. **(a)** Neuronal (left panel): in pIT, we found that the sum of the neuronal response to the face parts presented individually (blue) exceeded the response to the same parts presented simultaneously (i.e. a whole face, red). Each line is the mean response of 33 pIT sites normalized by the peak response to the whole face. Model (right panel): The magnitude of errors between stage 1 and stage 2 of the model showed a similar degree of sublinear integration (responses are normalized by peak response to the simultaneous input condition). **(b)** We extended the model to include two units in the third, output stage that responded to two learned input patterns (see **SI Methods**), increasing separation of patterns *A* and *B* in this high-level feature space (red markers in far right panel; 50 draws were made from distributions for *A* and *B* and were compared to relatively arbitrary inputs, blue markers). When alternating the two learned or familiar patterns *A* and *B*, activations in the top layer of the networks experienced greater

changes than when two randomly selected patterns were alternated (compare distances traversed in high-level feature space by red lines versus blue lines, far right panel). Because of this large change in high-level activations, transient errors were generated back through the network for learned patterns. Strong oscillations could be observed in the error signals between stage 1 and stage 2 of the model for alternated familiar inputs *A* and *B* (120 ms period) (red curve, middle right panel). In contrast, alternating novel inputs with similar amplitude but with patterns not matching the learned weights led to small amplitude oscillations in upstream error signals (blue curve). Note that the average response strength of state signals to novel and familiar inputs was matched at all network levels by construction. The differing response dynamics of error signals under familiar versus novel patterns are qualitatively consistent with the IT findings for novel versus familiar images (left, reproduced with permission from *Figs. 1 & 2b* of Meyer et al., 2014). **(c)** The average response to the whole face for pIT sites recorded 0 to 1 mm below the pial surface (left panel; superficial recordings, blue curve) and for sites 1 to 2.5 mm beneath the pial surface (red curve). (right panel) Average response of state units (blue) and error units (red) in stage 1 of the model. Note the lagged response in state units which is similar to the lagged response of units in superficial recordings (red curve, left panel).

Figure 7

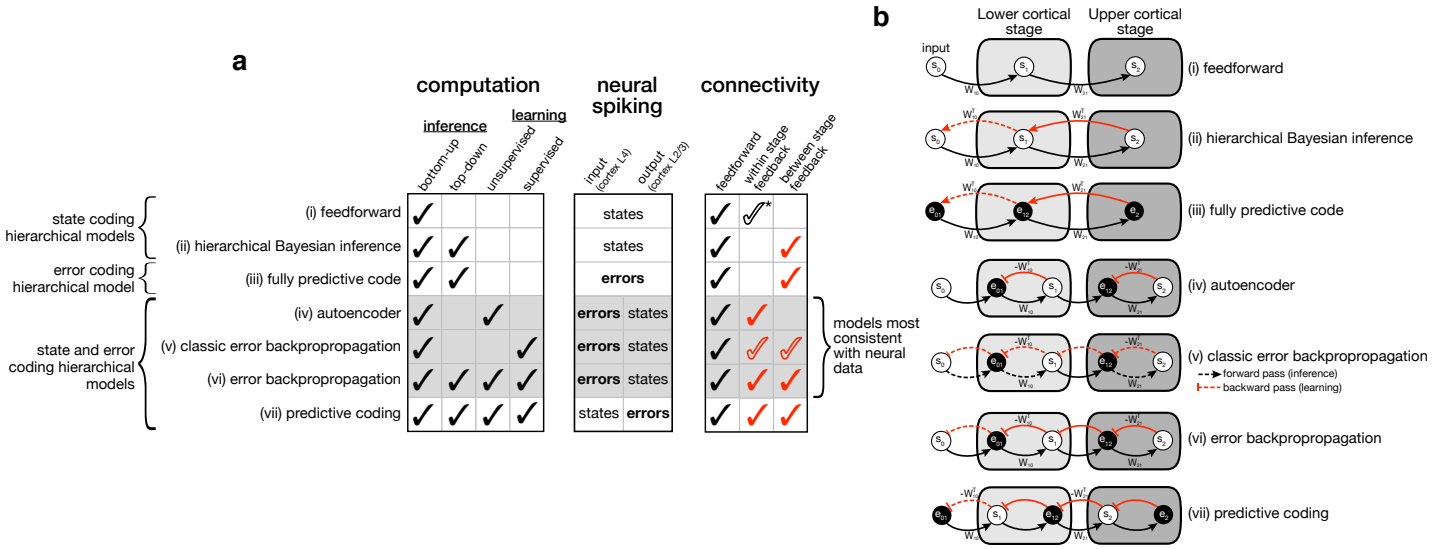
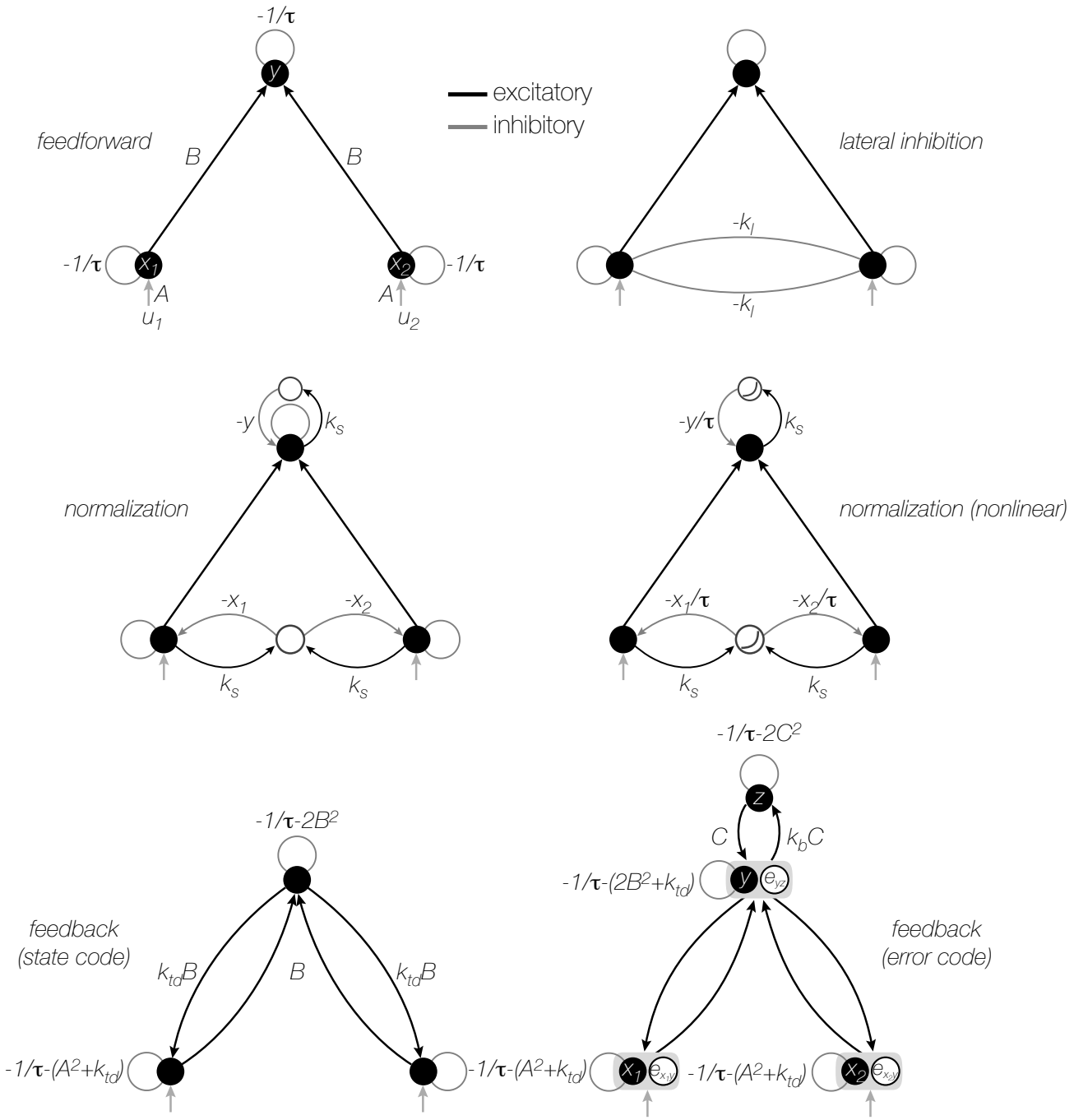


Figure 7 Summary of hierarchical model families. **(a)** Neural network models can be divided based on the computations that they perform (left table) and how they instantiate those computations through neurons and their connectivity (right two tables). (i): All model families shown build from the basic feedforward hierarchical model family to include either top-down inference, unsupervised learning, or supervised learning ((*) indicates that feedforward models can in some cases include local, within-stage recurrent connections such as the normalization models tested in **Fig. 5**). (ii): In hierarchical Bayesian inference (14)(31), feature estimates (i.e. inferred states of variables in the world) are communicated between processing stages to combine bottom-up estimates and top-down predictions. (iii): In a fully predictive code, neural spiking encodes the residual errors instead of directly encoding the feature estimates. This coding strategy amounts to a simple change of variables (see **SI Methods**) and is energy efficient in that the transmitted error signals are smaller (minimized by virtue of minimizing reconstruction cost) than state signals. A fully predictive coding model is shown here only for completeness as one example of an error coding model that performs efficient inference but not efficient learning, but see (vii) for a variant of predictive coding that also performs efficient learning. (iv): A simple form of a stacked autoencoder (33)(42) uses reconstructive error signals to drive unsupervised learning within each processing stage. (v): In classic error backpropagation (5)(19), errors are passed backwards between network levels during supervised training offline. (vi) Contemporary backpropagation approaches (42)(43) utilize error signals during online inference (similar to hierarchical Bayesian inference) through between-stage connections as well as during offline learning, through the same between-stage connections. Placing error coding units in the feedforward path makes these operations possible through basic neural integration and Hebbian plasticity mechanisms. (vii): Predictive coding (21)(40) is computationally similar to (vi), but it differs at the implementation level by specifically positing that error information (rather than state information) is passed to the next higher cortical stage. **(b)** Between-stage and within-stage connectivity diagrams corresponding to the models in **(a)**. Between-stage errors (black circles), measuring reconstruction performance, are computed in a similar fashion across models and can drive efficient hierarchical learning when coupled with state signals (white circles) (bottom four networks). The state and error computing networks only differ in the details of how error signals and state signals interact during inference and learning. Our data provide evidence for this large family of error-computing networks and rule out pure state-estimating models and variants including normalization and lateral inhibition (**Fig. 5**). The present data do not distinguish between the autoencoder and error backpropagation classes when directly compared (**Supplementary Fig. 2**); however, the stronger presence of state-like signals in the superficial cortical layers (**Fig. 6c**) argues against the predictive coding models in (iii) and (vii).

Supplementary Figure 1

a

Two-stage models (network diagrams)



b Two-stage models (connection matrices)

$$\dot{\mathbf{v}} = W\mathbf{v} + A\mathbf{u}, \;\; \mathbf{v} = \begin{bmatrix} x_1 \\ x_2 \\ y \end{bmatrix}, \;\; \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ 0 \end{bmatrix}$$

$$W_{feedforward} = \begin{pmatrix} -1/\tau & 0 & 0 \\ 0 & -1/\tau & 0 \\ B & B & -1/\tau \end{pmatrix}$$

$$W_{lateral\ inhibition} = \begin{pmatrix} -1/\tau & -k_l & 0 \\ -k_l & -1/\tau & 0 \\ B & B & -1/\tau \end{pmatrix}$$

$$W_{normalization} = \begin{pmatrix} -1/\tau - k_s \sum x & 0 & 0 \\ 0 & -1/\tau - k_s \sum x & 0 \\ B & B & -1/\tau - k_s y \end{pmatrix}$$

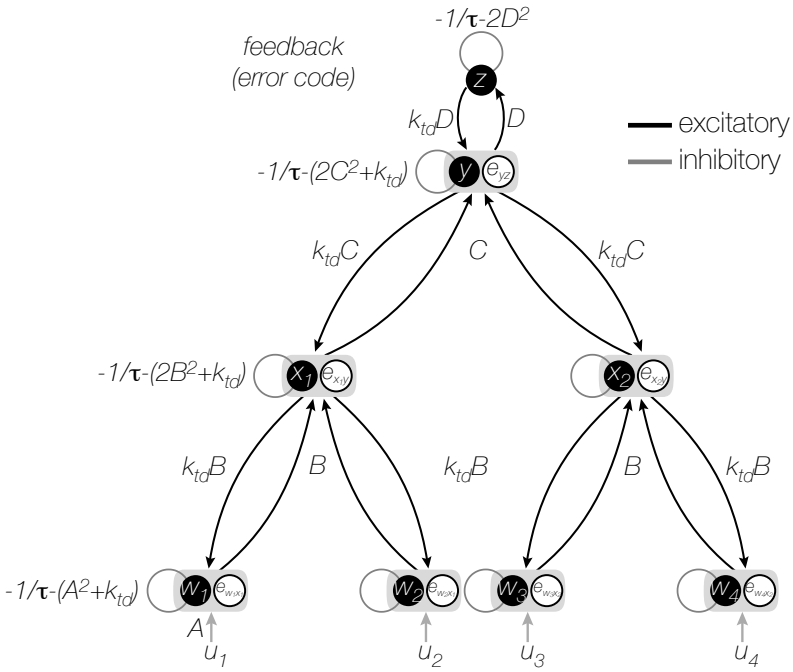
$$W_{normalization\ (nonlinear)} = \begin{pmatrix} -\dfrac{1}{\tau\sqrt{1 - k_s \sum x}} & 0 & 0 \\ 0 & -\dfrac{1}{\tau\sqrt{1 - k_s \sum x}} & 0 \\ B & B & -\dfrac{1}{\tau\sqrt{1 - k_s y}} \end{pmatrix}$$

$$W_{hierarchical\ generative\ (state)} = \begin{pmatrix} -1/\tau - (A^2 + k_b) & 0 & k_b B \\ 0 & -1/\tau - (A^2 + k_b) & k_b B \\ B & B & -1/\tau - 2B^2 \end{pmatrix}$$

$$W_{hierarchical\ generative\ (error)} = \begin{pmatrix} -1/\tau - (A^2 + k_b) & 0 & k_b B & 0 \\ 0 & -1/\tau - (A^2 + k_b) & k_b B & 0 \\ B & B & -1/\tau - (2B^2 + k_b) & k_b C \\ 0 & 0 & C & -1/\tau - 2C^2 \end{pmatrix}$$

c

Three-stage feedback model (network diagram)



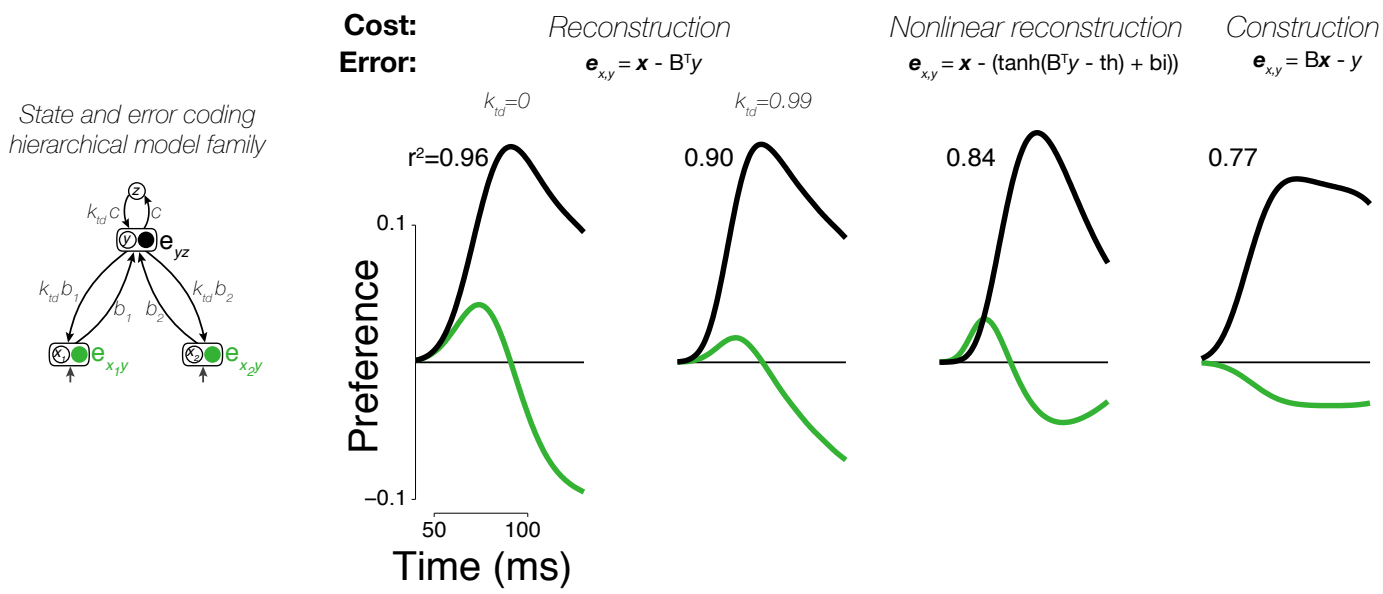
Three-stage feedback model (connection matrix)

$$W_{hierarchical\ generative\ (error)} = \begin{pmatrix} -1/\tau-(A^2+k_b) & 0 & 0 & 0 & k_bB & 0 & 0 & 0 \\ 0 & -1/\tau-(A^2+k_b) & 0 & 0 & k_bB & 0 & 0 & 0 \\ 0 & 0 & -1/\tau-(A^2+k_b) & 0 & 0 & k_bB & 0 & 0 \\ 0 & 0 & 0 & -1/\tau-(A^2+k_b) & 0 & k_bB & 0 & 0 \\ B & B & 0 & 0 & -1/\tau-(2B^2+k_b) & 0 & k_bC & 0 \\ 0 & 0 & B & B & 0 & -1/\tau-(2B^2+k_b) & k_bC & 0 \\ 0 & 0 & 0 & 0 & C & C & -1/\tau-(2C^2+k_b) & k_bD \\ 0 & 0 & 0 & 0 & 0 & 0 & D & -1/\tau-2D^2 \end{pmatrix}$$

Supplementary Figure 1 Model diagrams and equations. **(a)** Network diagrams of two-stage models. All models have two inputs (u_1, u_2), at least two hidden units (x_1, x_2) (a.k.a. layer 1 of the network), and at least one output unit (y). For simulations, the inputs u_1, u_2 are set independently to simulate each hidden node receiving different amounts of external drive, depending on the choice of the applied image relative to the unit's preferred image. The connection weights B = [b_1, b_2] transforming the hidden stage activations to the output unit are modeled as the same ($b_1 = b_2$). All units have self-connections that determine the degree of leak current set by the time constant τ. In the normalization models, the leak term is additionally controlled (linearly or nonlinearly) by the total activity in each stage (second row). In the feedback-based model, the feedback connections are symmetric to the feedforward connections with weights $B^T = [b_1, b_2]^T$, a column vector (third row). The error coding feedback model (bottom, right) has an additional stage that contributes to computation of error in the second stage (see **Methods** for details). **(b)** Equivalent system of differential equations for the networks shown in **(a)**. The general linear matrix equation governing the dynamics of the models is shown in the top row followed by the weight matrices W for each model in subsequent rows. The weight matrices determine the connection architecture of the models. Each model class uses a similar number of parameters to specify a different arrangement of connections. Strictly speaking, the two normalization model classes cannot

be written as linear systems of differential equations since the coefficients on the diagonal depend on the the state variables themselves creating nonlinear (multiplicative) terms, and for all simulations, solutions were computed using a standard iterative differential equation solver. **(c)** Extension of the two-stage model architectures to three stages for the feedback model (compare to two-stage diagram in **(a)**). The three-stage model has an additional hidden processing stage compared to the two-stage model. An extra node is introduced at the top of the hierarchy to produce an error signal in the third stage. The connection matrix for this model is shown below its network diagram (compare to matrix in **(b)**).

Supplementary Figure 2



Supplementary Figure 2 Variants of error coding hierarchical models using different algorithms for online inference. Existing forms of error computing networks can be distinguished by the type of online inference algorithm that they use (**Figure 7**). In one case, inference does not utilize top-down information between stages (autoencoder, classic error backpropagation; model classes (iv) and (v) in **Figure 7**). On the other hand, between-stage feedback can be used such as in more general forms of error backpropagation and predictive coding (models (vi) and (vii) in **Figure 7**). We approximated these two extremes by including a parameter (k_{td}, see **Methods**) controlling the relative weighting of bottom-up (feedforward) and top-down (feedback) evidence during online inference (first and second panels). We found that top-down inference between stages was not necessary to produce the appropriate error signals, and k_{td} was equal to zero in our best fitting two-layer and three-layers models (first panel is same model as two-layer error coding model in main text **Figure 5**) although models with $k_{td} \sim 1$ also performed well (second panel). Models can also differ in their goal (cost function) which directly impacts the error signals required (top row). Under a nonlinear reconstruction goal (emulating the nonlinear nature of spiking output), the resulting error signals are still consistent with our data (third column). A simple sigmoidal nonlinearity, however, did lead to additional details present in our neural data such as a rapid return of stimulus preference to zero in the hidden layer. When we tested a discriminative, construction goal more consistent with a supervised learning setting (e.g. classic error backpropagation) where bottom-up responses simply have to match a downstream target signal in classification tasks, we found that the errors of construction did not match the data as well as reconstruction errors (compare fourth column to first three columns; note absence of a reversal of selectivity in the hidden processing stage in fourth column) although both types of error outperformed all state-based models and were overall very similar (compare to **Fig. 5**).