Report

Neuron

High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy

Highlights

- Unexpected faces elicit prediction error responses in macaque face patch ML
- ML prediction errors exhibit identity specificity and view invariance
- Thus, ML prediction errors differ from ML tuning
- Instead, ML prediction errors share tuning of higher-order face patches AL/AM

Authors

Caspar M. Schwiedrzik,
Winrich A. Freiwald

Correspondence

c.schwiedrzik@eni-g.de (C.M.S.),
wfreiwald@rockefeller.edu (W.A.F.)

In Brief

Schwiedrzik and Freiwald show that macaque face patch ML computes prediction errors to unexpected faces. These signals exhibit tuning properties of higher-order face patches. This supports theories that suggest that lower brain areas compare bottom-up inputs with top-down expectations.

Schwiedrzik & Freiwald, 2017, Neuron *96*, 89–97
September 27, 2017
http://dx.doi.org/10.1016/j.neuron.2017.09.007

CellPress

Neuron
Report

High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy

Caspar M. Schwiedrzik[1,2,3,4,5,*] and Winrich A. Freiwald[1,*]
[1]Laboratory of Neural Systems, The Rockefeller University, New York, NY 10065, USA
[2]Neural Circuits and Cognition Lab, European Neuroscience Institute, 37077 Göttingen, Germany
[3]University Medical Center Goettingen, 37075 Göttingen, Germany
[4]Cognitive Neuroscience Laboratory, German Primate Center, 37077 Göttingen, Germany
[5]Lead Contact
*Correspondence: c.schwiedrzik@eni-g.de (C.M.S.), wfreiwald@rockefeller.edu (W.A.F.)
http://dx.doi.org/10.1016/j.neuron.2017.09.007

SUMMARY

Theories like predictive coding propose that lower-order brain areas compare their inputs to predictions derived from higher-order representations and signal their deviation as a prediction error. Here, we investigate whether the macaque face-processing system, a three-level hierarchy in the ventral stream, employs such a coding strategy. We show that after statistical learning of specific face sequences, the lower-level face area ML computes the deviation of actual from predicted stimuli. But these signals do not reflect the tuning characteristic of ML. Rather, they exhibit identity specificity and view invariance, the tuning properties of higher-level face areas AL and AM. Thus, learning appears to endow lower-level areas with the capability to test predictions at a higher level of abstraction than what is afforded by the feedforward sweep. These results provide evidence for computational architectures like predictive coding and suggest a new quality of functional organization of information-processing hierarchies beyond pure feedforward schemes.

INTRODUCTION

Cortical computations using spikes are expensive (Lennie, 2003). This suggests that the brain should employ an efficient coding strategy. One way to optimize information processing is to predict incoming stimuli based on past experience (Hawkins and Blakeslee, 2004). For example, a mechanism causing an expected stimulus to elicit a weaker response than the same stimulus when unexpected would reduce redundant information, and thus metabolic cost. How could this be implemented? Theoretical models, such as predictive coding (PC) (Friston, 2009; Huang and Rao, 2011; Rao and Ballard, 1999), often assume that expectations are formed by higher-level brain areas and are fed back to earlier ones (Figure 1A). When incoming sensory information and prediction mismatch, a prediction error (PE) is generated. Signaling the mismatch is considered more efficient than transmitting the entire bottom-up signal.

Importantly, PEs should be highly informative about the nature of the prediction signal, which is otherwise hard to measure: PC theory implies that PEs, although generated and measured in a lower-level area, reflect the tuning properties of cells that generate the prediction, i.e., of higher-level areas. This would constitute a major departure from the strict modularity of architectures relying solely on feedforward processing (Parkhi et al., 2015; Serre et al., 2007; Yamins and DiCarlo, 2016), where each stage performs a discreet operation on its inputs without external contributions.

An information-processing hierarchy especially well suited to put PC theory to the test is the macaque face-processing system. This system, residing in object-selective inferotemporal (IT) cortex, is a network of tightly interconnected face-selective areas (Grimaldi et al., 2016; Moeller et al., 2008; Tsao et al., 2006), each with a unique functional specialization (Freiwald and Tsao, 2010), implementing a three-level processing hierarchy (Figure 1B). Unique within IT cortex, the system's relevant stimulus class is known (faces) and also within-class selectivity for variations along two dimensions, head orientation and identity: cells in area ML are strongly tuned to head orientation ("view specificity"), profile-selective cells in AL respond to left and right profiles equally ("mirror-symmetric tuning"), and cells in AM are only weakly tuned to head orientation ("view invariance"). While tuning to head orientation decreases from ML via AL to AM, selectivity for facial identity across head orientations increases (Freiwald and Tsao, 2010). Because representations at the three processing stages differ not only quantitatively, but qualitatively, the face-processing system offers a unique opportunity to test PC theory: if the system utilizes predictions, if predictions are sufficiently detailed to differentiate between individual faces, and if prediction signals are fed back through the hierarchy, PEs in ML should not reflect local tuning properties, but those of AL and AM, the areas with the strongest feedback connections to ML (Grimaldi et al., 2016; Moeller et al., 2008). PEs in ML would then be more selective for identity than head orientation and possibly even reflect the mirror symmetry of AL.

Here, we investigate whether the face-processing system utilizes predictions, whether these predictions possess the granularity to differentiate between physically similar face stimuli, and whether PEs reflect properties of higher-level representations.

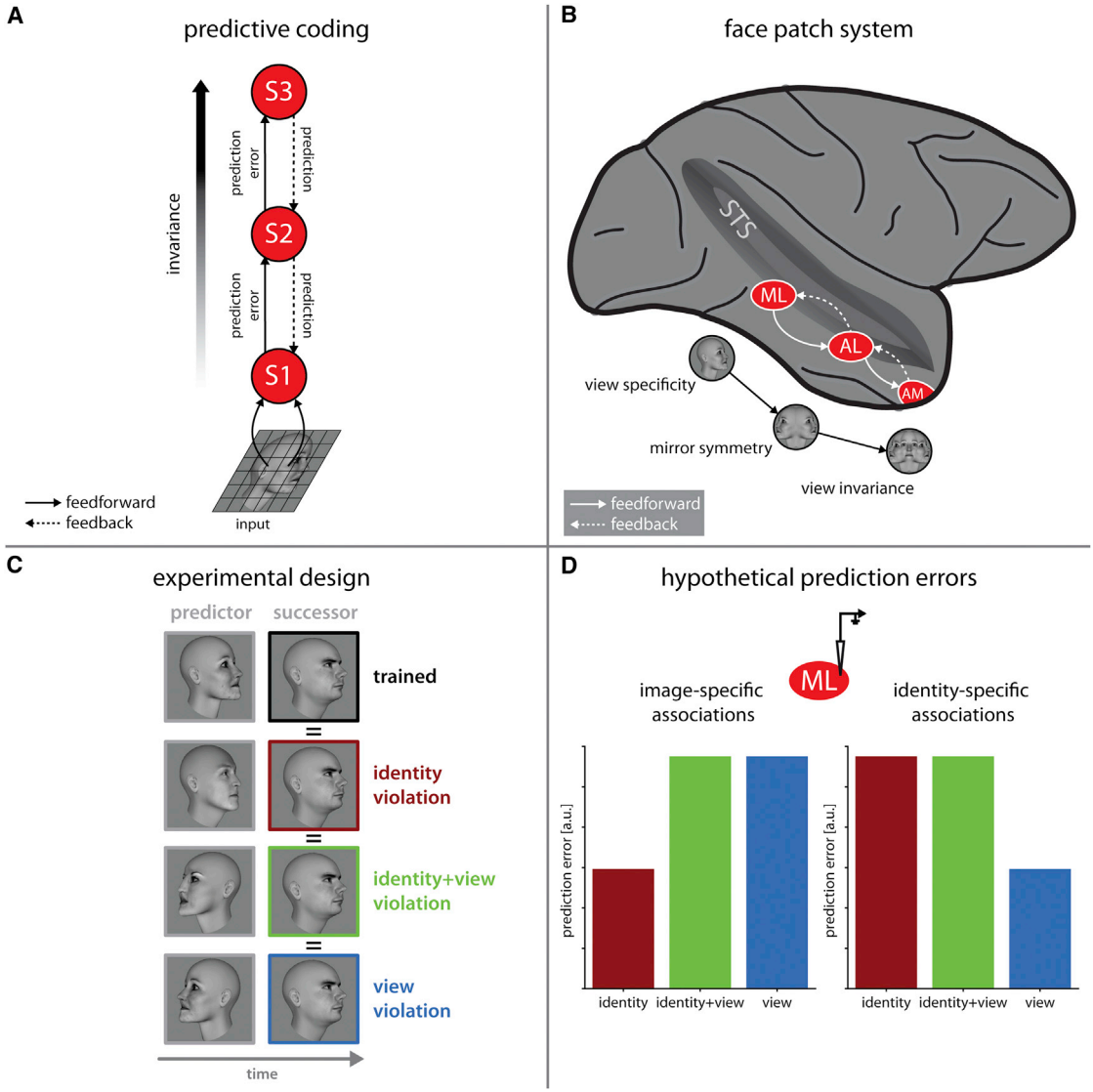

Figure 1. Background, Conditions, and Hypotheses

(A) Predictive coding proposes that higher processing stages send predictions to lower stages where they are compared with actual inputs. Any mismatch between prediction and input is signaled as a prediction error (PE).

(B) Subsequent stages in the face-patch system extract increasingly abstract information about facial identity by discarding information about head orientation. Neurons in face patch ML are view tuned, profile-selective AL neurons respond equally to left and right profile views, and AM represents identity largely independently of view. All stages are directly and reciprocally connected through feedback and feedforward connections.

(C) During the test phase of the paradigm, we presented trained pairs (black, 60% of all trials) and, critically, pairs with recombined head orientations and/or identities of the training partners such that predictors differed from the original pairing in view (blue), identity (red), or identity+view (green). Importantly, the second stimulus, the successor, remained identical across conditions. This allowed for a clean measurement of PEs as the response difference between unexpected and expected (but physically identical) successors.

(D) Depending on the source of the prediction signal, different patterns of PEs are expected. Image-specific PEs (left) should be maximal for deviations in head orientation, because changes in head orientation cause the largest pixelwise differences between images. In contrast (right), if learning generalizes across views to generate identity-specific predictions, PEs should be maximal for unexpected identities.

Good predictions are based on experience. To create those experiences, we employed an unsupervised statistical learning paradigm (see STAR Methods; Turk-Browne, 2012): for 4 weeks, we exposed monkeys to nine pairs of sequentially presented faces (Figure S1) to create specific associations such that, after learning, the appearance of the first stimulus (the predictor) forecast the appearance of a specific second one (the successor) (Meyer and Olson, 2011).

To determine the presence and properties of PEs, in the subsequent testing phase, we presented the trained face pairs as well as pairs that violated the learned associations along the two major tuning dimensions known to be encoded differentially

along the three stages of the face-processing system: head orientation and identity (Freiwald and Tsao, 2010). To elicit PEs in these dimensions, we created violation conditions by recombining predictors and successors into new pairs such that a successor could appear unexpectedly (in 40% of trials) after a predictor that differed from the successor's original training partner in identity, view, or both. We then contrasted responses of ML neurons to unexpected and expected but physically identical successor stimuli across the trained and three violation conditions (Figure 1C; STAR Methods). By manipulating only predictor stimuli while keeping successors physically identical (Meyer and Olson, 2011), we could isolate contextual effects of predictions from responses to successors (mitigating the confounding effects of stimulus selectivity that can arise when instead manipulating the successor). Comparing responses in the different violation conditions allowed us to test whether PEs in view-tuned ML reflect the properties of higher-level representations, identity specificity and view invariance (Figure 1D), as PC theory predicts.

RESULTS

We localized face patch ML with whole-brain fMRI in two rhesus monkeys (Figure 2A) and then targeted ML with electrophysiological recordings. For each neuron, we first assessed face selectivity and responsivity to the trained stimuli (see STAR Methods). Of 198 face-selective neurons, 80 were responsive to the trained stimuli (42 monkey M, 38 monkey Y). We then compared responses to the successor stimulus preceded by the trained face predictor with responses to the same stimulus but preceded by a different predictor diverging from the original predictor in view, identity, or identity+view. A majority of ML neurons emitted PEs, i.e., significantly stronger activity in the violation than in the trained conditions (Figure 2B): in 64% of responsive cells (24/42 in monkey M, 27/38 in monkey Y), PEs of varying size were elicited in response to violated expectations on the level of identity, view, or their combinations (Figures 2C and 2D). Effects were similar across animals (Figure S2), and the number of modulated cells was significant in both (all $p_{NPC} < 0.004$); hence, we report combined results. The pattern of firing rate reductions for expected relative to unexpected successors was not merely due to neural adaptation, because correlations between peak firing rates to consecutive predictor and successor stimuli were generally positive, not negative, as one would expect for adaptation, and did not differ between conditions (across neurons: $\chi^2 = 0.57$, $p = 0.7$; across trials per neuron: $\chi^2 = 1.95$, $p = 0.58$). Thus, the face-processing system can learn to actively generate predictions based on associations, even for the physically very similar stimuli in our training set (Figure S1), and about two-thirds of ML neurons compare these predictions against actual inputs, resulting in PEs.

What is the content of these predictive signals? ML could test predictions in different ways: as a general deviance detector, where PEs would signify contextual novelty for any unexpected stimulus, irrespective of the neurons' intrinsic sensitivity to the predicted face, or as an expectation-specific deviance detector, where PEs would be carried by neurons most informative about the predicted stimulus, a theoretical concept known as "precision" (Friston, 2009). To differentiate between these possibilities, we compared PEs for preferred and non-preferred successor stimuli across cells (Figures 3A and 3B). PEs were only elicited for preferred stimuli, resulting in a highly significant interaction between stimulus preference and predictability (Figure 3B; early: $p = 0.0032$, late: $p = 0.0078$). This shows that PEs reflect the precision-weighted testing of *specific* predictions about upcoming stimuli and thus that the face-processing system learns to generate predictions with sufficient granularity to distinguish variation within category, i.e., among similar face stimuli.

We then went on to test just how specific these predictions are: they could be based on associations on the level of head orientation, resulting in image-specific PEs (Figure 1D), as earlier research in anterior IT might suggest (Meyer and Olson, 2011). Alternatively, predictions could be identity specific, discarding deviations in head orientation that do not affect identity information—similar to tuning in AM (Freiwald and Tsao, 2010). To address this question, we separately investigated three kinds of prediction violations: identity, view, and identity+view. Each of these conditions elicited significant PEs with a mean latency of ~130 ms (Figures 3C–3G), averaging at 17% signal increase over the trained condition. A central claim of many PC models is that PEs signaled by a low-level area are successively suppressed by high-level feedback reconciling them with the prediction (Friston, 2009; Moreno-Bote and Drugowitsch, 2015; Murray et al., 2004). Thus, the influence of higher-order representations should increase with time. To test this prediction, we determined the time courses of PEs. All three violation conditions resulted in a significant PE during the early, transient response (Figure 3F). However, during the late, sustained phase of the response, view violation PEs first diminished and then completely vanished (time × condition, $p = 0.0002$), while PEs to identity violations and to identity+view violations remained significant and even slightly increased in strength (Figure 3G). This pattern of late responses is consistent with a source of the prediction signal that is identity, but not view, specific. These are the properties of representations in AL and AM. However, this signature might also arise locally, if ML neurons lost their tuning over time. But this was not the case: ML tuning to head orientation in the absence of predictions was strong and stable during stimulation, all the while head orientation information in the PE declined steadily and then disappeared (Figures 3H and 3I). Thus, invariance properties of higher-level representations emerge in ML PEs in time if faces appear in a highly predictable context.

An additional signature of the source of predictions are correlations between PEs and tuning: if the source of predictions is local, PEs are expected to depend on local tuning, but not when the source is a remote one. We found such correlations in the early, but not the late, response phase: early PEs were larger for more narrowly tuned cells (53–107 ms after peak response; Figure S3A), and early *view* violation PEs were larger for more narrowly *view*-tuned cells (63–188 ms after peak response; Figure S3B), while no such correlation existed for *identity* tuning and PEs for identity violations (Figure S3C). These correlations of local tuning and PEs suggest an early phase of PEs in ML that provides detailed information about the prediction violation and exhibits similar properties as local ML tuning (Freiwald and Tsao, 2010). None of the correlations between PEs and

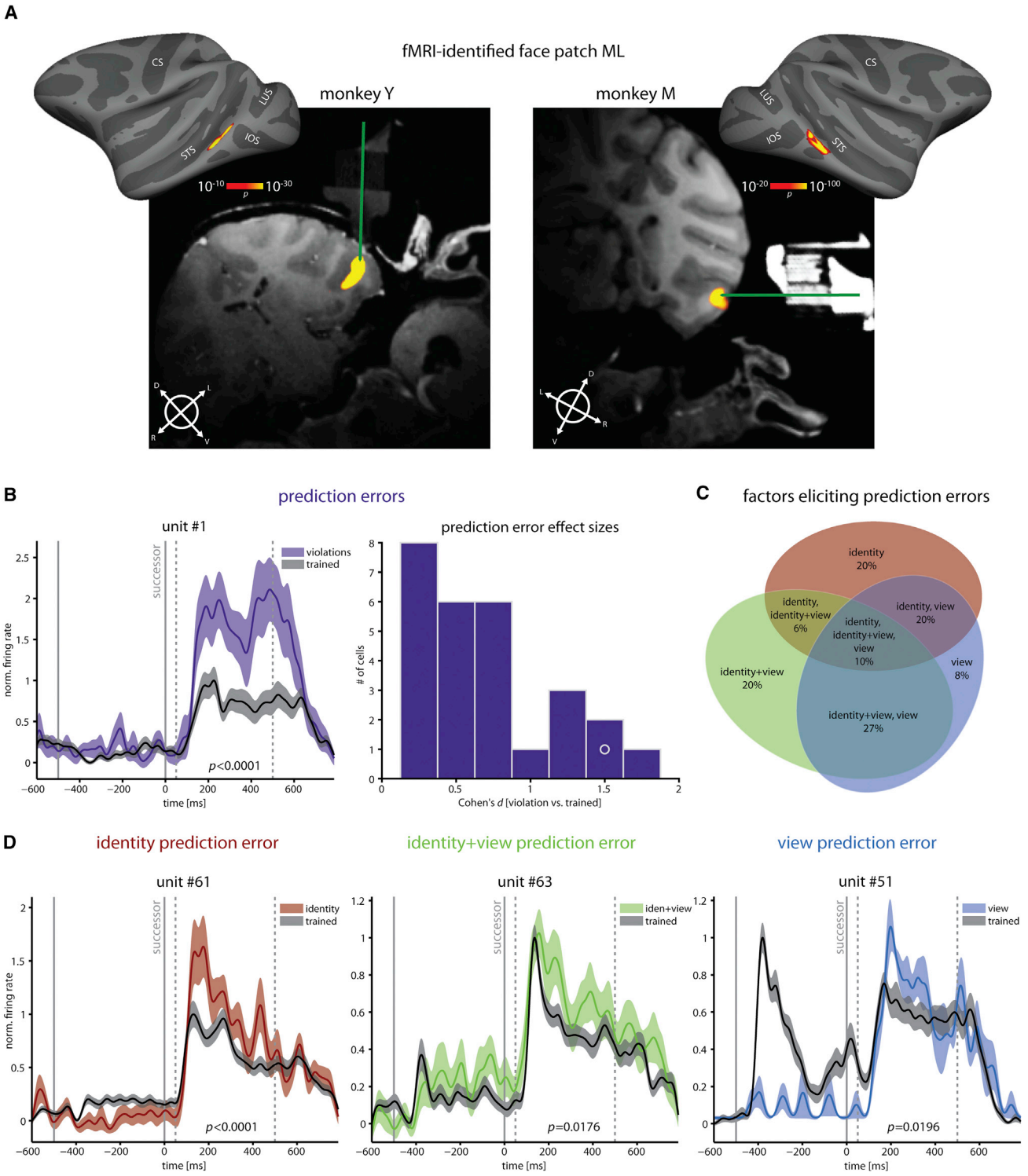

Figure 2. Single-Cell Prediction Errors

(A) fMRI-identified face patch ML on coronal slices and a surface template. Color coding represents significance of the contrast (faces versus objects and bodies), masked to highlight ML. Green lines indicate recording trajectories.

(B) Example single-cell prediction error (PE) and distribution of statistical effect sizes (average 50–500 ms) for neurons that showed PEs like the example cell (n = 27). The circle marks the example unit in the distribution of effect sizes.

(C) Percentage of responsive neurons showing statistically significant PEs (n = 51) to violations of identity, view, or combinations thereof.

(legend continued on next page)

ML tuning properties found in the early phase remained significant in the late phase (Figure S3). This indicates a time-dependent reduction of the impact of local tuning on PEs which parallels the emergence of view invariance and identity specificity of PEs in ML (Figures 3C and 3E).

Taken together, these results show that late PEs in ML do not reflect image-specific predictions. Late PEs, rather, are more identity than view specific, the reverse of the pattern of ML tuning. The gradual loss of sensitivity to view violations suggests that higher, view-invariant representations successively suppress PEs for view violations in ML, while PEs for identity are not diminished, thus preserving information about unexpected identities for further processing.

Identity selectivity and view tolerance of PEs in ML are properties shared with tuning in higher-level face areas AL and AM (Freiwald and Tsao, 2010). If area AL, the next higher-level processing stage directly and reciprocally connected to ML (Grimaldi et al., 2016; Moeller et al., 2008), was—partly or entirely—the source of predictions in ML, this suggests yet another functional signature for ML PEs (Figure 3J). AL is special among face patches in exhibiting mirror-symmetry confusion in profile-selective cells (Freiwald and Tsao, 2010); profile-selective neurons in AL, but not in ML (Figure 3K) and much less in AM, respond equally to mirror-symmetric profiles of the same face, but not to front views. Thus, predictions generated in AL should carry this mirror-symmetry signature as well. When we compared PEs in ML as a function of head orientation, we found that mirror-symmetric view violations elicited lower PEs in ML during the late phase of the response than non-mirror-symmetric view violations (Figure 3L). The timing of this effect overlaps with the peak of mirror-symmetric identity tuning in AL (Freiwald and Tsao, 2010). This suggests that during the late, sustained phase of PEs, higher-order representations specifically resembling those typically found in AL are involved in suppressing PEs about faces.

The effects of predictability that we observed occurred at the outset of the face processing hierarchy, but do they have an effect on face perception? If the neural effects described so far indeed transpired into behavior, they would predict a pattern of performance that differentiates identity- from view-based predictions. We tested behavior in humans, utilizing a slightly modified behavioral paradigm designed to tap into perceptual effects of implicit predictions (Figure 4A). 13 participants underwent training with the same stimuli as the monkeys, reduced to a single session of 20 min. Since predictability and feedback are thought to impact perception of signals embedded in high noise (Hupé et al., 1998; Summerfield and de Lange, 2014), we then tested how predictions affect detection of degraded faces. On each trial of this face priming paradigm, the predictor face served as a prime and the successor was degraded by phase scrambling. Unbeknownst to the subjects, we recombined successors and predictors to manipulate predictions on the level of identity, view, and view+identity on a subset of trials. Subjects detected faces in trained conditions faster (median difference in reaction time 327 ms, Figure 4B; Figure S4A) and more accurately (median difference in d' 0.13, Figure 4C; Figure S4B) than in untrained conditions, showing that learned predictability improves face detection. Accuracy was lowered in identity violation conditions ($p = 0.03$) but not in the view violation condition ($p = 0.39$; Figure 4C; Figure S4B). Thus, behavioral face prediction automatically generalizes across views, but not identity, paralleling PEs in ML (Figure 3G).

DISCUSSION

Our results provide direct evidence for learning in the adult macaque face-patch network, a system that supports a perceptual capacity with strong genetic determinants (Wilmer et al., 2010). The system, pre-wired after years of face processing, was capable of acquiring selectivity for arbitrary, artificial face pair sequences after a mere 30 days of passive exposure for a few hours per day in monkeys, and behavioral learning effects were already present after only 20 min of learning in humans. The face-processing system thus exhibits a remarkable potential to acquire entirely new environmental statistics. Even more so, stimuli were not associated as mere pictures, as earlier research in anterior IT might have suggested (Meyer and Olson, 2011), but were interpreted, automatically, as exemplars of specific facial identities. Nothing in the training regime had pre-empted this outcome, which supports, in a rather unexpected way, the conjecture that the face patch hierarchy's computational goal is the automatic extraction of view-invariant facial identity (Freiwald and Tsao, 2010).

The effects reported here also provide direct evidence for central tenets of PC theory: we can show that a lower-level area tests predictions about upcoming stimuli and that the resulting PEs are generated by many neurons, not just a few, which implement, through precision weighting, a computationally and metabolically efficient coding regime. Predictions prompted response modulations of ~17%, similar to contextual (Poort et al., 2016) and attentional (Treue and Martínez Trujillo, 1999) modulations in early and mid-level visual areas. Given the brief training, the arbitrariness of the associations, and the physical similarity of the stimuli, these likely are but a weak reflection of the full effect of predictions on information processing within this network. Our results, however, also constrain PC models in new ways: PEs undergo (at least) two phases, a hitherto undescribed early phase characterized by local tuning and hence possibly generated by locally computed predictions, and a later phase in which PEs resemble higher-level representations, as PC suggests. The finding that, during this later phase, higher-level properties emerge in PEs of an earlier area highlights an oft-overlooked consequence of the PC framework, the endowment of early processing stages with the capability to signal PEs that prescind from physical detail.

(D) Example PEs to identity (red, left), identity+view (green, middle), and view (blue, right) violations. Solid gray lines indicate stimulus onsets, and dotted lines are the period in which significance was assessed. p values reflect the result of a non-parametric combination (NPC) test against the null hypothesis that there is no effect in this time period. Firing rates are normalized to the min/max response in the trained condition. Shading around response average indicates ±SEM. Also see Figure S2.

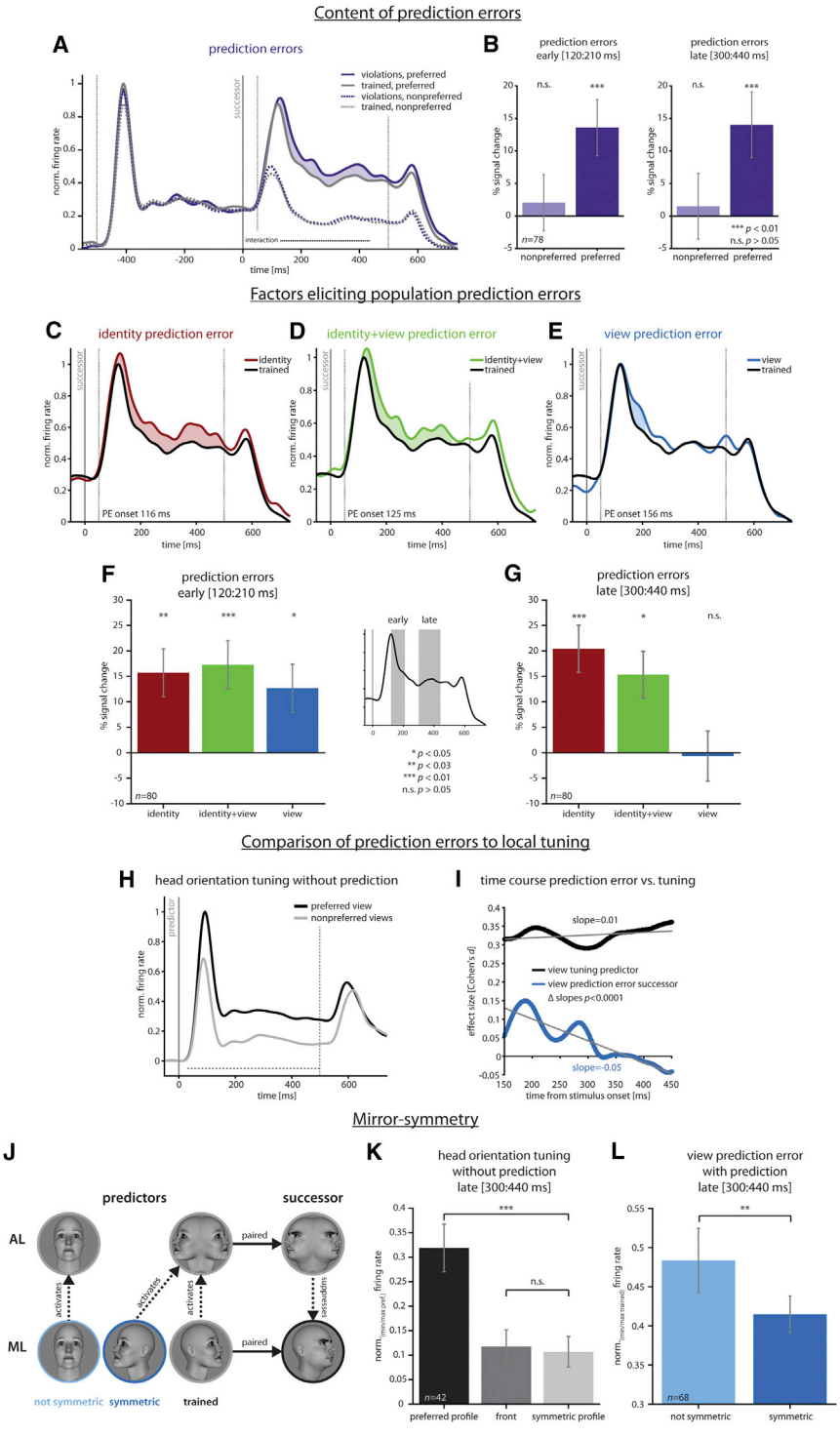

Figure 3. Properties of Prediction Errors

(A) Average responses to trained (gray) and violation (purple; identity, identity+view) conditions for preferred (solid lines) and non-preferred (dotted lines) successor stimuli across cells. Shading marks significant PEs (violations > trained), and the dotted horizontal line marks a significant condition (trained, violation) × preference interaction (corrected for multiple comparisons).

(B) PEs as percent signal change from the trained condition for preferred (dark purple) versus non-preferred (light purple) stimuli during the early, transient response (120–210 ms) and the late, sustained response (300–440 ms).

(C–E) PEs to identity (C; red), identity+view (D; green), and view (E; blue). Shading marks when violation conditions differ from the trained condition (corrected for multiple comparisons).

(F and G) PEs as percent signal change from the trained condition during the early (120–210 ms) and late (300–440 ms) response. View PEs were only significant (shading) during the former, not the latter, period (time × condition, $p = 0.0002$).

(F) Head orientation tuning of predictor stimuli. Predictor stimuli, which were not predictable, displayed a sustained difference for preferred and mirror-symmetric head orientations (corrected for multiple comparisons), exhibiting typical ML tuning. (G) While for stimuli lacking a predictive context (predictors) the difference between head orientations (black) was stable across time, view PEs for stimuli within a predictive context (successors, blue) decreased with time, resulting in view invariance (difference in slopes $p < 0.0001$).

(H) Schematic of hypothetical prediction formation between ML and AL. In ML and AL, learning links the two images in a pair. In ML, this link is view specific because ML neurons are view selective. In AL, the link is mirror symmetric because AL neurons exhibit mirror symmetry. Predictions from AL reduce PEs to the successor face in ML through feedback. Because AL pools ML inputs for left and right profiles, trained faces and mirror-symmetric view violations (dark blue) both predict successor faces, although the latter association was not explicitly trained. Non-mirror-symmetric view violations (light blue) are not pooled by the same AL neurons and do not reduce PEs.

(I) Responses of profile-preferring ML cells to stimuli lacking predictive context (predictors) are sharply view tuned, not mirror symmetric (300–440 ms).

(J) Mirror-symmetric (dark blue) view violations elicit smaller PEs than non-mirror-symmetric (light blue) view violations in the late phase of the response to the successor (300–440 ms). Error bars in (B), (F), (G), (K), and (L) indicate SEM, corrected for between-cell variability. Firing rates in (A), (C)–(E), (H), (K), and (L) are normalized to the min/max response in the (preferred) trained condition/to the preferred head orientation. Dotted vertical lines in (A), (C)–(E), and (H) bound the window for multiple comparison correction. Slopes in (I) were obtained through a linear least-squares fit. Also see Figure S3.

PC proposes that this transfer of properties results from recurrent processes in which higher areas pass predictions to lower areas. Our results suggest that high-level face representations impact information processing in a lower-level face area. This offers stimulus prediction as a functional role for the face-processing system's abundant feedback projections (Grimaldi et al., 2016; Moeller et al., 2008), which had so far remained enigmatic (DiCarlo et al., 2012; Freiwald and Tsao,

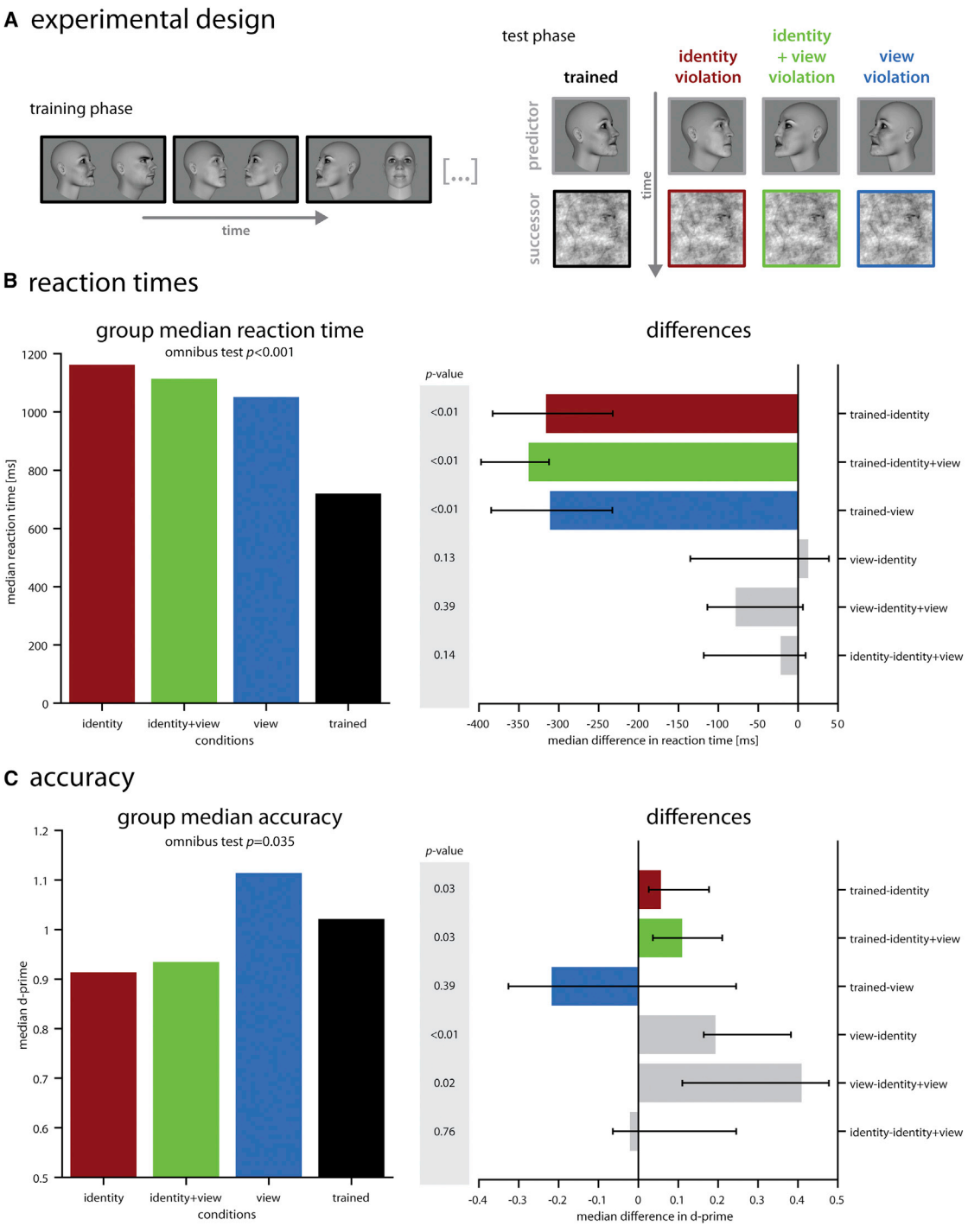

Figure 4. Behavioral Learning Effects on Face Detection

(A) Human subjects were trained with the same face pairs as the monkeys for 20 min. Subsequently, we tested for learning and generalization in a face priming task. We presented trained pairs (black), view (blue), identity (red), and identity+view (green) violations. The predictor image served as the prime. Subjects decided whether the successor was a face embedded in noise or only noise.

(B) Median reaction times. The left shows median reaction times per condition, and the right shows median differences between conditions on which statistics were performed.

(C) Median accuracy (measured by d′) and differences between conditions. (B and C) Omnibus tests simultaneously compare all conditions; p values for planned comparisons are given in gray boxes. Error bars indicate bootstrapped 95% confidence intervals. Also see Figure S4.

2010; Meyers et al., 2015). The gradual development of view-invariant PEs in ML indicates that beyond bidirectional conduction delays, time-consuming, possibly iterative computations need to be carried out before PEs are suppressed and higher and lower areas converge on an internally consistent account of the environment. This likely includes the full emergence of view-invariant tuning in AL/AM, which peaks at around 300 ms (Freiwald and Tsao, 2010). Finally, the functional precision of effects (differentiating individual faces) implies that feedback connections, similar to feedforward projections, are wired with high specificity even within the confines of a face area only a few millimeters in diameter. Thus, our results suggest a new quality of functional organization of information-processing hierarchies in IT cortex beyond currently predominant feedforward schemes, whereby top-down predictions alter online information processing.

Disregarding head orientation information in predictions comes at the expense of losing information about physical image properties in PEs. This computational strategy may bolster the extraction of identity information by emphasizing processing at a higher level of abstraction across the face-processing hierarchy. However, PC theory also affords updating predictions if the discrepancy between predictions and actual inputs is too large (Friston, 2009). This keeps internal models from which predictions are derived in check. The concurrent emergence of view invariance and decreasing impact of local tuning properties on the PE we found speak for an increasing role of abstract, higher-order representations in suppressing PEs. This does not rule out that if violations had diverged more strongly from the trained pairs, e.g., along a dimension such as species, new predictions could have been learned. The impact of such PEs on higher-order representations in AL/AM is an interesting target for future studies.

Taken together, we show how incidental learning of statistical regularities interacts with the organizational principles of cortical hierarchies to allow the brain to optimize processing resources and to generalize from image-specific predictions to abstract rules (Aslin and Newport, 2012), revealing deeper "knowledge" than the mere association of two specific images already at early stages of processing. This exemplifies the profound penetration of perception by experience and suggests a new information-processing quality for object-recognition hierarchies.

STAR★METHODS

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

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Monkey Subjects
  - Human Subjects
- METHOD DETAILS
  - Monkey Experiments
  - Human Experiments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Monkey Experiments
  - Human Experiments
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2017.09.007.

AUTHOR CONTRIBUTIONS

C.M.S. and W.A.F. designed the experiments, interpreted data, and wrote the manuscript. C.M.S. collected and analyzed data.

ACKNOWLEDGMENTS

We thank L. Melloni, C. Olson, C. Schroeder, A. Winkler, and W. Zarco for guidance and support; D. Sonnenberg for help with human data acquisition; and S. Rasmussen, L. Diaz, and A. Gonzalez for veterinary and technical care. This work was supported by a Human Frontier Science Program Long-Term Fellowship (LT001118/2012-L, C.M.S.), an Irma T. Hirschl/Monique Weill-Caulier Trusts Award (W.A.F.), a Pew Scholar Award in the Biomedical Sciences (W.A.F.), a McKnight Scholars Award (W.A.F.), the New York Stem Cell Foundation (W.A.F.), the National Eye Institute (R01 EY021594, W.A.F.), the National Center for Advancing Translational Sciences (CTSA UL1 TR001866, RUCCTS), the NSF Science and Technology Center for Brains, Minds, and Machines (CCF-1231216/5710003506, W.A.F.), and the National Science Foundation (INSPIRE Track 2 DBI-1343174, W.A.F.). W.A.F. is a New York Stem Cell Foundation-Robertson Investigator. C.M.S. is currently funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 706519. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Received: April 28, 2017
Revised: July 6, 2017
Accepted: September 7, 2017
Published: September 27, 2017

REFERENCES

Aslin, R.N., and Newport, E.L. (2012). Statistical learning: from acquiring specific items to forming general rules. Curr. Dir. Psychol. Sci. *21*, 170–176.

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

Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.

Fisher, C., and Freiwald, W.A. (2015). Contrasting specializations for facial motion within the macaque face-processing system. Curr. Biol. *25*, 261–266.

Freedman, D., and Lane, D. (1983). A nonstochastic interpretation of reported significance levels. J. Bus. Econ. Stat. *1*, 292–298.

Freiwald, W.A., and Tsao, D.Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science *330*, 845–851.

Friston, K. (2009). The free-energy principle: a rough guide to the brain? Trends Cogn. Sci. *13*, 293–301.

Grimaldi, P., Saleem, K.S., and Tsao, D. (2016). Anatomical connections of the functionally defined "face patches" in the macaque monkey. Neuron *90*, 1325–1342.

Hautus, M.J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. Behav. Res. Methods Instrum. Comput. *27*, 46–51.

Hawkins, J., and Blakeslee, S. (2004). On Intelligence (Times Books).

Huang, Y., and Rao, R.P. (2011). Predictive coding. Wiley Interdiscip. Rev. Cogn. Sci. *2*, 580–593.

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

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. Percept. Psychophys. *49*, 227–229.

Lancaster, T., and Jae Jun, S. (2010). Bayesian quantile regression methods. J. Appl. Econ. *25*, 287–307.

Lennie, P. (2003). The cost of cortical computation. Curr. Biol. *13*, 493–497.

Meyer, T., and Olson, C.R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. Proc. Natl. Acad. Sci. USA *108*, 19401–19406.

Meyers, E.M., Borzello, M., Freiwald, W.A., and 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*, 7069–7081.

Micallef, L., and Rodgers, P. (2014). eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses. PLoS ONE *9*, e101717.

Moeller, S., Freiwald, W.A., and Tsao, D.Y. (2008). Patches with links: a unified system for processing faces in the macaque temporal lobe. Science *320*, 1355–1359.

Moreno-Bote, R., and Drugowitsch, J. (2015). Causal inference and explaining away in a spiking network. Sci. Rep. *5*, 17531.

Murray, S.O., Schrater, P., and Kersten, D. (2004). Perceptual grouping and the interactions between visual cortical areas. Neural Netw. *17*, 695–705.

Ohayon, S., and Tsao, D.Y. (2012). MR-guided stereotactic navigation. J. Neurosci. Methods *204*, 389–397.

Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In Proceedings of the British Machine Vision Conference (BMVC), *1*, 6.

Pesarin, F. (2001). Multivariate Permutation Tests with Applications in Biostatistics (John Wiley & Sons).

Poort, J., Self, M.W., van Vugt, B., Malkki, H., and Roelfsema, P.R. (2016). Texture segregation causes early figure enhancement and later ground suppression in areas v1 and v4 of visual cortex. Cereb. Cortex *26*, 3964–3976.

Quiroga, R.Q., Nadasdy, Z., and Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. Neural Comput. *16*, 1661–1687.

Raghunathan, T. (2003). An approximate test for homogeneity of correlated correlation coefficients. Qual. Quant. *37*, 99–110.

Rao, R.P., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.

Rolls, E.T., and Tovee, M.J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. J. Neurophysiol. *73*, 713–726.

Salimi-Khorshidi, G., Smith, S.M., and Nichols, T.E. (2011). Adjusting the effect of nonstationarity in cluster-based and TFCE inference. Neuroimage *54*, 2006–2019.

Samonds, J.M., Potetz, B.R., and Lee, T.S. (2014). Sample skewness as a statistical measurement of neuronal tuning sharpness. Neural Comput. *26*, 860–906.

Schwiedrzik, C.M., Zarco, W., Everling, S., and Freiwald, W.A. (2015). Face patch resting state networks link face processing to social cognition. PLoS Biol. *13*, e1002245.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., and Poggio, T. (2007). A quantitative theory of immediate visual recognition. Prog. Brain Res. *165*, 33–56.

Smith, S.M., and Nichols, T.E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage *44*, 83–98.

Summerfield, C., and de Lange, F.P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nat. Rev. Neurosci. *15*, 745–756.

Tippett, L.H.C. (1931). The Methods of Statistics (Williams & Norgate).

Treue, S., and Martínez Trujillo, J.C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. Nature *399*, 575–579.

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

Turk-Browne, N.B. (2012). Statistical learning and its consequences. Nebr. Symp. Motiv. *59*, 117–146.

Wang, P., and Nikolić, D. (2011). An lcd monitor with sufficiently precise timing for research in vision. Front. Hum. Neurosci. *5*, 85.

Wilcox, R. (2012). Introduction to Robust Estimation and Hypothesis Testing, Third Edition (Academic Press).

Willenbockel, V., Sadr, J., Fiset, D., Horne, G.O., Gosselin, F., and Tanaka, J.W. (2010). Controlling low-level image properties: the SHINE toolbox. Behav. Res. Methods *42*, 671–684.

Wilmer, J.B., Germine, L., Chabris, C.F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., and Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. Proc. Natl. Acad. Sci. USA *107*, 5238–5241.

Winkler, A.M., Ridgway, G.R., Webster, M.A., Smith, S.M., and Nichols, T.E. (2014). Permutation inference for the general linear model. Neuroimage *92*, 381–397.

Winkler, A.M., Webster, M.A., Brooks, J.C., Tracey, I., Smith, S.M., and Nichols, T.E. (2016). Non-parametric combination and related permutation tests for neuroimaging. Hum. Brain Mapp. *37*, 1486–1511.

Yamins, D.L., and DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. *19*, 356–365.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
Ferumuxytol (Feraheme)	AMAG Pharmaceuticals	NDC: 59338-775-0
Deposited Data		
Electrophysiological and psychophysical data	This paper	http://dx.doi.org/10.6084/m9.figshare.5126326
Software and Algorithms		
eulerAPE	Micallef and Rodgers, 2014	http://www.eulerdiagrams.org/eulerAPE/
Freesurfer	Fischl, 2012	RRID: SCR_001847
(MATLAB) Permutation Analysis of Linear Models (PALM) toolbox	Winkler et al., 2016	https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/PALM/
(MATLAB) Planner toolbox	Ohayon and Tsao, 2012	http://tsaolab.caltech.edu/?q=Planner
(R) rmdzero function	Wilcox, 2012	https://dornsife.usc.edu/labs/rwilcox/software/
(MATLAB) Spectrum, Histogram, and Intensity Normalization and Equalization (SHINE) toolbox	Willenbockel et al., 2010	http://www.mapageweb.umontreal.ca/gosselif/SHINE/
(MATLAB) wave_clus toolbox	Quiroga et al., 2004	http://www2.le.ac.uk/centres/csn/research-2/spike-sorting/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Caspar M. Schwiedrzik (c.schwiedrzik@eni-g.de).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Monkey Subjects

All animal procedures met the National Institutes of Health *Guide for Care and Use of Laboratory Animals*, and were approved by the local Institutional Animal Care and Use Committees of The Rockefeller University (protocol number 12585-H) and Weill-Cornell Medical College (protocol number 2010-0029), where magnetic resonance imaging (MRI) was performed. Data were acquired in 2 male, pair-housed adult macaque monkeys (*Macaca mulatta*, 9 (monkey M) and 8.7 (monkey Y) kg, age 6 (monkey M) and 5 (monkey Y) years).

Human Subjects

All human procedures were approved by the Institutional Review Board of The Rockefeller University (protocol number WFR-0741). Data were acquired in 22 subjects (10 female, 1 left handed, mean age 33.4 years) after they gave written informed consent. No sample size estimate was performed, but sample size was selected based on previous studies. All subjects reported normal or corrected-to-normal vision and no history of neurological and/or psychiatric disease.

METHOD DETAILS

Monkey Experiments

Surgery

Implantation of MR-compatible headposts (Ultem; General Electric Plastics), recording chambers (Crist Instruments), ceramic screws (Rogue Research), and acrylic cement (Grip Cement, Caulk; Dentsply International, and Palacos, Heraeus Kulzer GmbH) followed standard anesthetic, aseptic, and postoperative treatment protocols (Moeller et al., 2008).

Magnetic Resonance Imaging

MRI data were acquired on a 3T scanner (Siemens TIM Trio). Functional data were acquired with an AC88 gradient insert (Siemens) and a custom 8-channel phased-array receive surface coil with a horizontally oriented single loop transmit coil (L. Wald, MGH/HST

Martinos Center for Biomedical Imaging) while the monkeys were in sphinx position. Before scanning, the contrast agent ferumoxytol (Feraheme, AMAG Pharmaceuticals; 8-10 mg of Fe per kg body weight) was injected into the femoral vein to increase the signal-to-noise ratio (SNR). To localize face areas, we acquired 16 (monkey Y) and 17 (monkey M) runs of functional (T_2*-weighted) gradient-echo echoplanar imaging (EPI). Each run consisted of 196 volumes of 54 horizontally oriented slices (field of view [FOV] 96 mm, voxel size 1 × 1 × 1 mm, repetition time [TR] = 2 s, echo time [TE] = 16 ms, echo spacing [ESP] = 0.63 ms, bandwidth [BW] = 1860 Hz/Px, flip angle [FA] = 80 degrees, no gap) acquired in interleaved order with phase partial Fourier 7/8, and two times generalized autocalibrating partially parallel acquisitions (GRAPPA) acceleration, covering the whole brain. Additionally, we obtained field maps which allowed subsequent EPI undistortion. Anatomical images were obtained in a separate session using a T_1-weighted magnetization-prepared rapid gradient echo (MPRAGE) sequence (FOV 128 mm, voxel size 0.5 × 0.5 × 0.5 mm, TR = 2.53 s, TE = 3.07 ms, ESP = 7.3 ms, BW = 190 Hz/Px, FA = 7 degrees, 240 slices) and a custom 1-channel receive coil (L. Wald, MGH/HST Martinos Center for Biomedical Imaging) while the monkeys were anesthetized (isoflurane 1.5%–2%) and positioned in an MR-compatible stereotactic frame (Kopf Instruments). Blood vessels were visualized using the contrast agent Gadolinium (0.2 mL per kg body weight).

To localize the face patch ML, we used a standard face localizer (Fisher and Freiwald, 2015). In short, subjects fixated on a white dot at the center of the screen while we presented images of human and/or monkey faces, human and/or monkey body parts and/or headless bodies, manmade objects, and fruits, intermixed with baseline periods in which only the fixation dot was shown in a block design (FOB from here on). Each block lasted 24-30 s. Fluid reward was delivered after variable periods of time (2–4 s) during which the subject maintained fixation within 2 degrees of the fixation dot. Only runs in which the subjects reached at least 90% fixation stability were used for analyses. Visual stimulation and reward were controlled using in house software (Visiko, M. Borisov). Stimuli were projected on a back-projection screen using a video projector (NEC NP3250, refresh rate 60Hz, resolution 1024 × 768 pixel) with a custom lens. Eye position was measured at 120 Hz using a commercial eye monitoring system (ISCAN).

Stimuli and Training

For the main experiments, we generated 36 3-dimensional human faces with a neutral expression and no hair in FaceGen (v3.5.3, Singular Inversions). For each face, we extracted five head orientations (0, 30, 60, 300, 330 degrees), resulting in a total of 180 images. Images were converted to black and white and luminance normalized using SHINE (Willenbockel et al., 2010). For training, we selected 18 unique images (6 per head orientation 0, 60, 300 degrees) from the full set, each showing a different face. These were joined into 9 pairs (Figure S1). Pairs were arranged such that one identity-view combination would uniquely predict one other identity-view combination, while assuring that head orientation was fully balanced across pairs (e.g., 3 different identities at 0 degrees head orientation were paired with 3 different identities at 0, 60 and 300 degrees head orientation, respectively).

In the training phase, each monkey then received extensive exposure (30 days in monkey M, 32 days in monkey Y) to the 9 pairs of face images in order to establish associations between stimuli, and, separately, to the full set of 180 faces (18 paired faces + 162 unpaired faces) to familiarize them with the stimuli later used to determine tuning for head orientation and identity.

For the face pair training, each trial started with a fixation period (500 – 2000 ms), followed by the first image in a pair (500 ms), followed by the second image in a pair (500 ms) with no ISI. Images (5 dva) were presented foveally on a gamma-corrected CRT monitor (NEC Multisync FE2111SB, refresh rate 100 Hz, resolution 1280 × 1024 pixel, viewing distance 57 cm) in a darkened booth. The sequence of pairs was arranged such that transition probabilities within pairs (i.e., between stimuli) were 100%, while transition probabilities between pairs (i.e., between trials) were at minimum and balanced across pairs. This was done to assure that monkeys only learned associations between the stimuli within pairs. Monkeys were rewarded with a drop of juice and/or water if they continuously fixated on a white, centrally presented fixation dot within a 3 × 3 dva window for 2-4 s. Thus, there was no systematic relationship between the occurrence of a pair on the screen and reward. In separate runs, monkeys were also exposed to the full set of 180 faces (18 paired faces + 162 unpaired faces), presented in random order and distinguished from pair training runs by stimulus timing (200 ms on, 200 ms off) and by using a blue fixation dot. These additional training runs mimicked the tuning measurements during later recordings. Eye position was monitored at 120 Hz using an ISCAN system. Visual stimulation and reward were controlled using in house software (Visiko, M. Borisov).

Electrophysiology

Chamber locations, grid angles, and electrode trajectories were planed based on functional and contrast-enhanced anatomical MRIs using Planner (Ohayon and Tsao, 2012). ML was targeted in the left hemisphere of monkey Y and the right hemisphere of monkey M. For recordings, tungsten electrodes (500k-3M Ohm, FHC) were back loaded into metal guide tubes, which also served as the reference. Guide tube length was set to reach the top of the dura. The electrode was slowly advanced using an oil hydraulic micromanipulator (Narishige Scientific Instrument). Neural signals were amplified and recorded at 30k Hz (Blackrock Microsystems). Spikes were detected online based on their waveform and later sorted offline using wave_clus (v2.5) (Quiroga et al., 2004).

On each recording day, the electrode was initially advanced based on MRI-planning until the border of face patch ML was reached. We then recorded from each single unit encountered along the electrode trajectory within ML. After a unit had been isolated, we ran four experiments, described in detail below. First, we mapped the size and location of the excitatory receptive field region by moving face stimuli on the screen. Second, after optimizing stimulus size and position, we determined category selectivity, using the FOB stimulus set (faces, objects and bodies) from the fMRI face localizer. Third, we assessed responsivity to the full face stimulus set; the latter experiment also served to determine identity and view tuning of the cell (IVT form here on). Fourth, we ran the main experiment in which we presented the trained pairs as well as untrained pairs of faces. The order of experiments was not randomized,

and the experimenter was not blinded to the experimental conditions. After completing all four experiments, we advanced the electrode until the next cell was reached. During all experiments, monkeys were required to fixate within a 3 × 3 dva fixation window and were given drops of juice and/or water as reward for fixation performance every 2 to 4 s.

Experimental Conditions

We adapted a paradigm by Meyer and Olson (2011) that is specifically aimed at eliciting PEs. A key feature of this paradigm is that responses to *physically identical* stimuli (successors) can be measured in different predictive contexts. Thus, the effect of predictions can be directly extracted without the need for assumptions about neural tuning properties, as is often necessary in other experimental designs on Predictive Coding. Specifically, during the main experiment, we presented the trained pairs, as well as pairs in which the expected sequence of stimuli within a pair was violated. To create the violation conditions, we systematically recombined the trained facial identities and head orientations of the first images in the trained pairs (the predictors) with the second images in the trained pairs (the successors) to generate novel pairs. This allowed us to test the influence of correct (trained) versus incorrect (violation) expectations on the processing of the second stimulus in the pair (the successor). We then measured PEs following their standard definition used in the literature as the difference between expected and unexpected successors. Crucially, by only manipulating the predictor and keeping the second stimulus identical across conditions, we could isolate contextual effects of expectations onto the successor, which could otherwise be masked by differences in the cells' responsivity to different successor stimuli. In the view violation conditions, we kept the identity of the first stimulus the same as in the trained condition, but changed the head orientation to one of the other two head orientations (e.g., if identity A was trained with 60 degrees, we now presented the same identity A with 0 and 300 degrees head orientation; Figure 1C; Figure S1). Hence, if predictions were based on associations that were formed on the level of head orientation, we would expect PEs in response to the successor in this condition. In the identity violation condition, we recombined the successor with predictors from other pairs that had the same head orientation as the trained predictor (e.g., if identity A had been trained at 60 degree head orientation, we now presented trained identities D and G at the same 60 degree head orientation as predictors, inducing a wrong expectation about the successor; Figures 1C; Figure S1). If predictions relied on associations on the level of identity, disregarding view, we would expect PEs in the 'identity' but not the 'view violation' condition, since the former preserves head orientation but not identity, while the latter preserves identity but not head orientation. In the identity+view violation condition, we recombined the successor with predictors from other pairs that had both a different identity and a different head orientation than the originally trained first image (e.g., if identity A had been trained at 60 degree head orientation, we now presented trained identities E/F/H/I at 0 and 300 degrees head orientation, again resulting in wrong expectations about the successor; Figure 1C; Figure S1). This condition allowed us to assess the relationship between the factors head orientation and identity in generating PEs.

Because size and position of the stimuli during testing were tailored to each cell's receptive field, while training occurred at a fixed size and location, any learning effect can be considered size- and position-invariant. To maintain the training effects, the trained/untrained ratio was maintained at 1.5 during testing. Monkeys completed minimally 270 trials per recorded unit, with an inter-trial-interval (ITI) ranging between 0.5 and 2 s. The order of conditions was randomized.

Identity/View Tuning Measurements (IVT)

To determine responsivity to the stimulus set and to assess identity and view tuning independently of the experimental conditions, we conducted a separate identity/view tuning measurement for each recorded cell (IVT). To this end, we presented the full set of all 180 trained and untrained faces in random order and rapid succession (200 ms on, 200 ms off) for minimally 1260 trials (total). As during initial training, the tuning measurements were distinguishable from the main experimental conditions both in stimulus timing and through an explicit cue, the color of the fixation dot (main experiment: white, tuning measurement: blue).

Human Experiments

Stimuli, Training, and Tasks

Stimuli were displayed on an LCD monitor (Samsung 2233RZ; Wang and Nikolić, 2011), resolution 1680 × 1050 pixel) at a refresh rate of 120 Hz. Subjects viewed the screen from a distance of 58 cm. The experiments were conducted in a darkened room. Constant head position was assured by the use of a chinrest with forehead support. Stimulus delivery and response collection were controlled using Presentation (v16.4, Neurobehavioral Systems).

Face stimuli were the same as in the monkey experiments. Additionally, we created noisy versions of the stimuli by parametrically combining each image with a phase-scrambled version of itself (1%–100% noise). During all experiments, stimuli were presented centrally at 10 × 10 dva, and subjects were instructed to fixate on a foveally presented fixation dot.

To individually determine the noise level at which subjects could detect faces (versus noise) at 75% accuracy, we first conducted a threshold measurement using the weighted up-down staircase method (Kaernbach, 1991). To this end, we presented on each trial one of 6 different faces (0, 60 and 300 degrees view angle) that were not part of the training set embedded into noise, or a noise-only image. The subject's task was to determine on each trial whether a face was present or not by means of a button press on a standard keyboard. The starting noise level was 80%. Whenever the subject responded correctly, the noise level was increased by 9 percent; whenever the subject responded incorrectly, the noise level was decreased by 3 percent. Stimulus duration was 50 ms and the ITI varied randomly between 500 and 1000 ms. Each subject completed 72 trials (50% faces, 50% noise only). The average threshold across subjects was 74.9% noise.

Next, subjects passively viewed the same training pairs as the monkeys over 6 blocks of 72 trials (a total of 432 trials). All stimulus parameters were the same as during the monkey training. As during monkey training, the transitional probabilities within a pair were fixed at 100% and were kept balanced and at minimum between pairs. The total duration of the training phase was approximately 20 min.

After the training phase, we conducted a priming experiment to assess whether learning face pairs affected face detection performance. As in the monkey experiments, we presented trained pairs and pairs in which we had recombined predictors with successors such that the predictors in the new pairs differed from the originally trained predictors in view, identity, or identity+view. Predictors were presented for 500 ms and acted as a prime. Right after the predictor, we showed the successor for 50 ms, embedded into the noise level previously determined to yield 75% accuracy in face detection, or a noise-only image, as target. The subject's task was again to determine whether a face was present in the second image, or not. To maintain the training effects, the trained/untrained ratio of the pairs was fixed at 1.667. Subjects completed 4 blocks of 144 trials (total 576 trials). The ITI varied randomly between 0.5 and 1.5 s. The order of conditions was randomized.

QUANTIFICATION AND STATISTICAL ANALYSIS

Monkey Experiments

Magnetic Resonance Imaging

MRI data were analyzed in Freesurfer (v5.1, http://surfer.nmr.mgh.harvard.edu/) (Fischl, 2012) and MATLAB (R2011b, The Mathworks). The first 5 volumes of each functional run were excluded to prevent T_1 saturation effects. Preprocessing included slice scan time correction, motion correction, and geometric distortion correction by means of a field map. Outliers in the time courses were detected semi-automatically based on a threshold of median absolute deviation (MAD) = 3.5 in the mean whole-brain time course and later excluded from analyses. We identified face patch ML following established procedures (Moeller et al., 2008): For each animal, we calculated a General Linear Model (GLM) with the stimulation conditions as predictors as well as six orthogonalized nuisance regressors accounting for motion artifacts. As in previous studies, ML was identified based on anatomical location and relative position (Moeller et al., 2008; Schwiedrzik et al., 2015) in unsmoothed, uncorrected significance maps (monkey Y $p < 10^{-67}$, monkey M $p < 10^{-119}$) resulting from the contrasts [faces versus objects and bodies]. MRI data from both monkeys were also used in a previous study (Schwiedrzik et al., 2015).

Electrophysiology

We recorded from a total of 198 neurons (106 monkey M, 92 monkey Y). The number of neurons was chosen to match or exceed that in similar studies. Spike density functions (SDF) were obtained by convolving spike trains with a Gaussian kernel (σ = 17 ms) and decimated to 1 kHz. Unless otherwise noted, only trials on which the monkey continuously fixated from 100 ms before stimulus onset until the end of the images sequence and on which second stimuli were shown to which a cell was responsive entered the main analyses. Each trial was baseline corrected by subtracting the average prestimulus baseline (tuning measurement 50 ms, main experiment 100 ms) from every time point. To determine whether a cell was responsive to a given stimulus, we tested for each cell whether there was a significant positive response to the stimulus ($p < 0.05$) within the first 150 ms post stimulus onset across trials, by means of a Wilcoxon signed rank test on the tuning measurement data. Using an independent dataset assured that the assessment of responsivity was independent of the experimental factors in the main experiment. Using this criterion, a total of 80 neurons (42 monkey M, 38 monkey Y) were found to be responsive to the experimental stimuli and entered the main analyses.

Data from the main experiment was analyzed with the General Linear Model (GLM), using a permutation framework (Freedman and Lane, 1983; Winkler et al., 2014) to assess significance in PALM (v0.94) (Winkler et al., 2016). We carried out two sets of analyses: first, we determined whether a cell's response to the successor stimuli was significantly modulated by manipulations of view and/or identity of the predictor stimuli, i.e., whether there was a contextual modulation, by entering the trial-by-trial data from each cell into a GLM with the orthogonal factors view (violation, trained) and identity (violation, trained). Second, we tested specifically for the presence of prediction errors (PEs), i.e., larger responses to unpredicted than predicted successor stimuli in individual cells by assessing simple effects in the GLM, i.e., by contrasting [view violation versus trained], [identity violation versus trained], [identity+view violation versus trained], and [view, identity, identity+view violations versus trained]. All analyses were carried out on each time point between 50 and 500 ms after the onset of the second stimulus. To determine whether a cell showed a significant effect, we used nonparametric combination (NPC) (Pesarin, 2001) with Tippett's combining function (Tippett, 1931): NPC is a method to perform joint inference on multiple data, in this case time points, with only minimal assumptions. We first tested each contrast at each time point separately, using permutations performed synchronously across time points. Synchronized permutations have the benefit that they account for any dependence among the partial tests. The test statistics t for each partial permutation test was then transformed into a pseudo p-value, and all pseudo p-values were combined using Tippett's combining function. This combining function uses the minimum pseudo p-value across tests as its test statistic and assesses the null hypothesis that the null hypotheses for all partial tests are true, and the alternative hypothesis that any is false, resulting in a single hypothesis test per cell. Because it can be formulated as the maximum statistic across tests, it can be used to control the familywise error rate (FWER) across time. To assess whether the number of cells showing significant PEs was itself significant, we again used NPC, combining the cell-wise NPC p-values. The Venn diagram in Figure 2C showing the resulting percentages of responsive neurons with statistically significant PEs was drawn using ellipses in eulerAPE to achieve accurate area-proportionality (Micallef and Rodgers, 2014).

Population level analyses were carried out using the permutation GLM after averaging trials per cell and condition. To correct for multiple comparisons in these time-resolved analyses, we used threshold-free cluster enhancement (TFCE) (Smith and Nichols, 2009), a cluster-based method to control the family-wise error rate which mitigates the arbitrary setting of a cluster-forming threshold. TFCE was applied separately to the first and the second stimuli in a pair to avoid the detrimental effects of nonstationarities on cluster-based inference (Salimi-Khorshidi et al., 2011), but this did not change the overall pattern of results. PE latencies were determined as the first time point at which the contrast [violation versus trained] reached statistical significance after correction for multiple comparisons. Percent signal change in the violation conditions was calculated relative to the trained condition. Mean percent signal change across conditions was obtained by averaging all significant time points, corrected for multiple comparisons, across the three violation conditions. The time-resolved analyses were complemented by averaging the time courses in an early (120-210 ms post stimulus onset) time window, capturing the transient response, and a late (300-440 ms post stimulus onset) time window capturing part of the sustained response in which the time point to time point rate of change had stabilized, and conducting permutation tests on these averages. Together, the single cell and population analyses allowed us to assess whether face patch ML was sensitive to prediction violations, which factors elicited PEs, and which time course PEs had.

In addition, we assessed the relationship between the cells' tuning properties and PEs. To this end, we computed each cell's view and identity tuning in the 200 ms following its onset latency (i.e., the first of more than 20 samples > 1 SD above baseline within the first 300 ms post stimulus) using the skewness of the distribution of average responses across views and identities, respectively, from the independent tuning measurement as a measure of tuning sharpness (Samonds et al., 2014). Higher skewness indicates sharper tuning. Additionally, we computed each cell's (lifetime) sparseness (Rolls and Tovee, 1995) as

$$\frac{\left(\sum_{i=1}^{N} R_i/N\right)^2}{\left(\sum_{i=1}^{N} R_i^2/N\right)}$$

where N equals the number of stimuli, with responses R_i not baseline subtracted (Freiwald and Tsao, 2010). Lower sparseness indicates that the cell responds to fewer images in the stimulus set. We then calculated the Spearman rank correlation between tuning/sparseness and the effect size of prediction errors, measured as the statistical effect size Cohen's *D* from the GLM, in sliding windows (window size 100 ms, step size 1 ms), followed by TFCE to control for multiple comparisons in time.

The role of stimulus-specific adaptation was assessed by correlating the average firing rates within a 50 ms time window centered on the grand average peak response to the first and second stimulus, respectively. Stimulus-specific adaptation would result in a negative correlation between these peak firing rates, and differences in these correlations between conditions would speak for a differential role of adaptation. Correlation analyses were computed on the single cell level (across trials) and on the population level (across neurons). Fisher-transformed Spearman correlation coefficients across trials were compared using a nonparametric Friedman ANOVA, and across neurons using a χ^2 test for dependent correlations (Raghunathan, 2003).

Human Experiments

Individual face detection thresholds were determined by averaging all but the first two reversals from the threshold experiments. Data from 9 subjects had to be excluded from analyses because they failed to follow task instructions (final n = 13, 5 female, 1 left handed, mean age 33.6 year). For the analysis of the main experiment, trials with reaction times shorter than 100 ms as well as trials on which reaction times exceeded a threshold of 3x the MAD were excluded from further analyses. We then calculated *d'* with the loglinear correction to avoid infinite *z*-scores (Hautus, 1995) as a bias-free measure of face detection accuracy, as well as the average reaction time for hits. For statistical analyses, we used a percentile bootstrap procedure comparing median differences between all conditions (function *rmdzero* in R, v3.2.3) (Wilcox, 2012), followed by planned comparisons. The same results were obtained with Bayesian statistics (Figure S4), using the substitution posterior for the median (Lancaster and Jae Jun, 2010).

DATA AND SOFTWARE AVAILABILITY

Data have been deposited in the Figshare repository (https://www.figshare.com) under accession number http://www.dx.doi.org/10.6084/m9.figshare.5126326.