Predictions enable top-down pattern separation in the macaque face-processing hierarchy

Tarana Nigam[1,2,3,4], Caspar M. Schwiedrzik[1,2,3*†]

[1.] Neural Circuits and Cognition Lab, European Neuroscience Institute Göttingen – A Joint Initiative of the University Medical Center Göttingen and the Max Planck Society, Grisebachstraße 5, 37077 Göttingen, Germany
[2.] Perception and Plasticity Group, German Primate Center – Leibniz Institute for Primate Research, Kellnerweg 4, 37077 Göttingen, Germany
[3.] Leibniz ScienceCampus 'Primate Cognition', Göttingen, Germany
[4.] International Max Planck Research School 'Neurosciences', Georg August University Göttingen, Grisebachstraße 5, 37077 Göttingen, Germany

* Correspondence: C.M.S. (c.schwiedrzik@eni-g.de)
† Lead contact

Summary

Distinguishing faces requires well distinguishable neural activity patterns. Contextual information may separate neural representations, leading to enhanced identity recognition. Here, we investigate how predictions derived from contextual information affect the separability of neural activity patterns in the macaque face-processing system, a 3-level processing hierarchy in ventral visual cortex. We find that in the presence of predictions, early stages of this hierarchy exhibit well separable and high-dimensional neural geometries otherwise characteristic only of the top of the hierarchy. This is accompanied by a systematic shift of tuning properties from higher to lower areas, endowing lower areas with higher-order, invariant representations instead of their feedforward tuning properties. Thus, top-down signals dynamically transform neural representations of faces into separable and high-dimensional neural geometries. Our results provide evidence how predictive context transforms flexible representational spaces to optimally use the computational resources provided by cortical processing hierarchies for better and faster distinction of facial identities.

Highlights

- Predictions increase pattern-separation and dimensionality in early face-processing areas.
- Early face-processing areas dynamically inherit tuning properties from higher areas.
- Predictions enable top-down separability and invariance for identity recognition.

Keywords

pattern separation, predictive coding, statistical learning, feedback, plasticity, face-processing, inferotemporal cortex, macaque monkey, functional magnetic resonance imaging

Introduction

We have all mistaken someone or been mistaken for somebody else. Indeed, recognizing people by their face is a significant computational challenge both for humans and machines, especially in light of variations in appearance, e.g., with a different hairstyle or pose. This is particularly evident when distinguishing lookalikes (or “doppelgänger”) like the actresses Natalie Portman and Keira Knightley in a movie (Fig. 1A)[1]. To do so, we need additional *contextual* information, e.g., information from previous scenes, which we use to make predictions[2]. Predictions render recognition better and faster, yet the computations and neural implementation mediating those effects remain poorly understood. In particular, it is unknown how contextual predictions optimize sensory neural representations, especially for distinguishing individuals by their face.

The reason why we struggle to distinguish Natalie from Keira is that similar stimuli lead to correlated neural activity patterns. To distinguish stimuli like faces, neural representations should have minimal overlap and hence be "separable". Separability is largely determined by stimulus properties, i.e., how physically similar faces are, and only achieved at hierarchically higher areas in the ventral visual stream, where neural representations are also robust against variations like pose change[3]. How can incorporating predictive information enable distinguishing identities? Predictive processing theories suggest that higher-order areas generate predictions that are communicated to lower areas via feedback pathways[4]. We hypothesized that incorporating higher-order predictive information leads to more separable (Fig. 1A) and invariant (Fig. 1B right) representations already in lower areas of the hierarchy.

To investigate how predictability optimizes neural representations for distinguishing identities, we turned to the macaque face-processing system (Fig. 1B left). This information processing hierarchy in inferotemporal (IT) cortex is highly specialized for faces, and the neural representations and transformations across successive stages of processing are well understood. In the ventral part of this system, representations become successively invariant to face view while the system extracts facial identity: the lowest area, ML, represents face shape and possesses view-specificity; the intermediate area AL has mirror-symmetric representations; at the apex of the hierarchy, in AM, representations are view-invariant and contains facial appearance information for identity recognition[5,6]. We used functional magnetic resonance imaging (fMRI) to access these qualitatively distinct representations in all three face-areas simultaneously. By training animals to expect specific faces based on temporally preceding information[7] (Fig. 1C), we could test how predictions affect neural representations, and whether feedback pathways are involved.

We find a gradient of pattern-separability along the face-processing hierarchy with the highest separability at the top, in AM, even when no predictive context exists. Once predictions come into play, pattern-separability increases in lower stages of the hierarchy. This is accompanied by the emergence of abstract, view-invariant representations at these lower processing stages: rather than the view-specific tuning properties they show in the absence of predictions, we find that lower face-areas express representations characteristic of higher stages of processing when stimuli are predictable. This suggests that high-level prediction signals cascading-down the face-processing system dynamically transform neural representational spaces, endowing them with increased pattern-separability and invariance. This may enable lower areas to abstract from irrelevant features for enhanced distinction of identities. Our findings provide a new take on the flexibility of neural representations in IT and suggest a specific mechanism, top-down pattern-separability, by which predictions exert their effects.

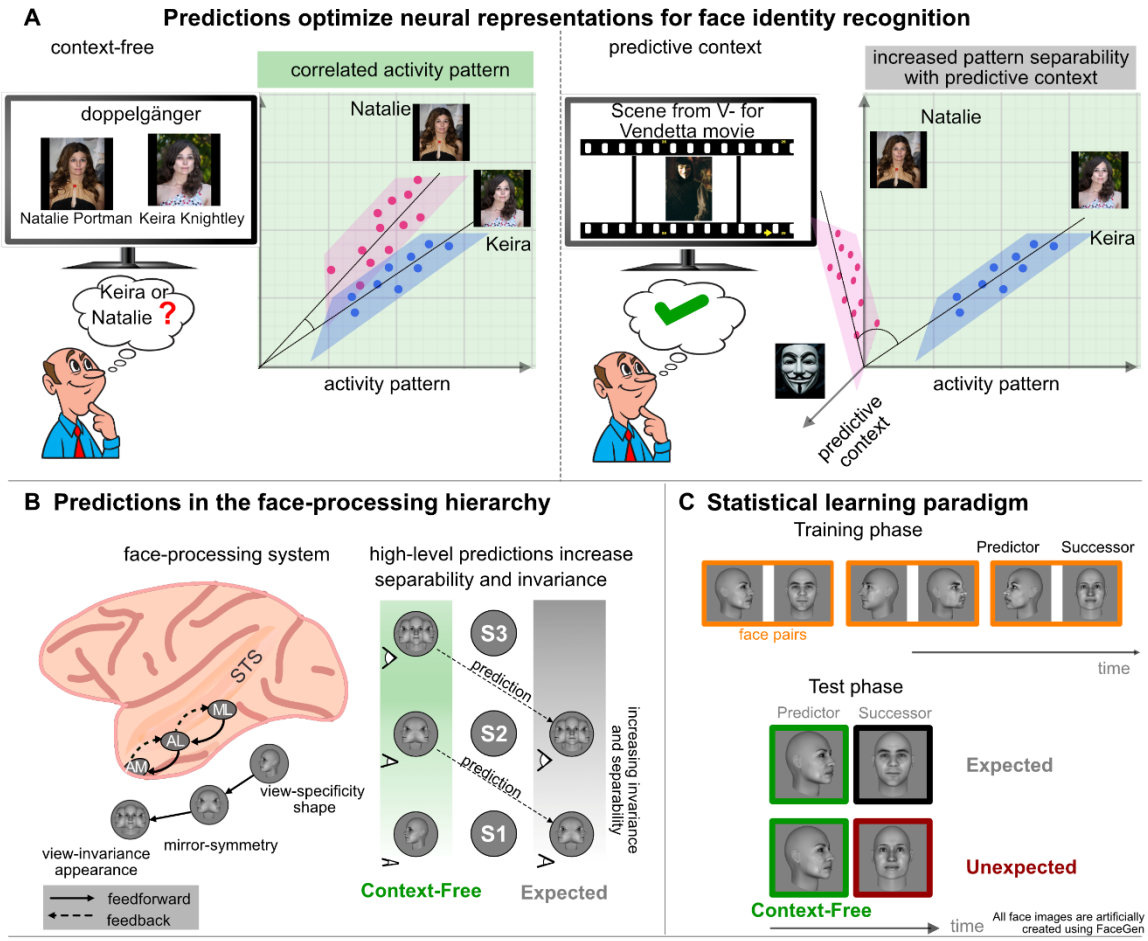

Figure 1: Increasing pattern separability through top-down predictions. *A.* Similar faces, like those of the lookalikes Natalie Portman and Keira Knightley, are hard to distinguish because they lead to poorly separable neural activity patterns. Contextual predictions, derived, e.g., from a previous scene, can aid distinguishing similar stimuli. We hypothesize that predictive context facilitates recognition by increasing separability between neural activity patterns. ***B.*** In the face processing system, the lowest area ML has view-specific shape representations, the intermediate area AL has mirror-symmetric representations, and the top of the hierarchy, AM, has view-invariant appearance representations. All stages are directly and

reciprocally connected. Higher face areas contain neural representations with minimal overlap (high separability) and view-invariance. We hypothesize that incorporating high-level predictive information (via feedback pathways) increases invariance and separability in lower areas, ***C.*** To create contextual predictions, we exposed monkeys to face pairs in a statistical learning paradigm. Once learned, the first face in a pair (the "predictor") predicts the second face in a pair (the "successor"). In the test phase, we presented predictable face pairs (60% of all trials) along with their violations during fMRI. Faces in A, B and C were computer-generated using Facegen. Additional images are licensed from Adobe, Alamy, and Shutterstock.

Results

As a first step in investigating whether and how predictions alter neural coding in the face-processing system, we localized three face-areas - ML, AL, and AM – in two macaque monkeys using a standard face localizer for whole-brain fMRI. We then trained the monkeys to associate temporally contiguous faces into pairs using statistical learning[7] (Fig. 1C). Once learned, the first face in a pair (the "predictor") fully predicts the second face (the "successor"). We exposed the monkeys to nine face-pairs for a minimum of three weeks, providing ample time to elicit effects of predictability in IT cortex[8]. To assess the acquisition of face-pair associations, we evaluated pupil dynamics, an established marker of statistical learning[9]. Indeed, we found pupil entrainment to the frequency of the face-pairs after training (Supplementary Fig. S1). Once we established face-pair learning in the monkeys, we investigated if and how coding spaces changed as a function of contextual predictions across the face-processing system using a rapid event-related design for fMRI (Fig. 1C). In the *expected* condition (60% of trials), we presented learned face-pairs, such that predictions were elicited once the predictor had been shown and thus took effect when the successor appeared. On violation trials, predictions went unfulfilled as we presented an *unexpected* identity in the expected view or an unexpected view of the expected identity as the successor[7]. Since our focus was on investigating how predictability contributes to the identity recognition, one of the main computational goals of face-processing, we mostly concentrate on conditions involving facial identity.

Before evaluating how predictive context optimizes neural representations for face individuation, we first established how neural separability for identity increases across the face-processing hierarchy in the absence of predictive context (the *context-free* condition). In each of the three face-areas we measured cosine distances between activity patterns elicited by different faces (Fig. 2Ai). We report average results because these distances were highly correlated across monkeys (Spearman's rho=0.95, p=3.5e-04). Pattern-separability increased along the hierarchy (ML<AL<AM, paired Hotelling test, all p<1.462e-11, corrected for multiple-comparisons). This corroborates previous electrophysiological studies in IT cortex outside the face-processing system[10] and indicates that, in the absence

of predictability, high levels of pattern-separability for facial identities are only achieved in AM, the top of the face-processing hierarchy.

After determining baseline pattern-separability across the face-processing system, we evaluated whether predictability aids face identification by increasing pattern-separability in lower areas. We compared separability across face identities between the expected and context-free conditions (Fig. 2Aii): when faces were expected, we found that lower areas in the face-processing hierarchy, ML and AL, increased pattern-separability for identities by ~10° from 52.1° to 61° in ML and 67.7° to 77.2° in AL (Fig. 2B; William-Watson test, ML: $F(1,70)=79.74$, $p=1.067e-12$; AL: $F(1,70)=34.49$; $p=2.619e-07$, corrected for multiple comparisons). In contrast, pattern-separability in the higher area AM remained unchanged (84.7° vs. 87.5°; $F(1,70)=3.53$; $p=0.0644$, corrected for multiple comparisons). This suggests that high separability otherwise characteristic of the top of the processing hierarchy can already be attained in lower areas if predictive information exists - potentially facilitating readout of information from well decorrelated representations.

Predictive processing theories suggest that predictions serve to compute prediction errors (PE) as a deviation between the prediction and the sensory input to reduce redundancy in information transmission[11]: redundant information is removed, optimizing the dynamic range of neurons and allowing them to efficiently signal decorrelated or unpredicted information[12]. In terms of pattern separation, this may entail higher pattern-separability when PEs (to unexpected faces) occur. To evaluate this possibility, we compared pattern separation between the violation and context-free conditions in the three face areas. Prediction violations increased pattern-separability in the entire face-processing hierarchy (Fig. 2B; William-Watson test, ML: 77.7° vs. 52.1°, $F(1,70)=485.84$, $p<0.0001$; AL: 82.8° vs. 67.7°, $F(1,70)=115.65$, $p=4.44e-16$; AM: 89.6° vs. 84.7°, $F(1,70)=14.00$, $p=0.0003$, corrected for multiple comparisons). AM patterns were, in fact, almost orthogonal in this condition (mean 89.6°, one-sample test for mean angle vs. 90°, $p>0.05$). Interestingly, prediction violations increased pattern-separability even beyond the separability achieved in predictable conditions, but only in lower stages of the hierarchy (paired Hotelling test, ML: 77.1° vs. 61°, $F(2,34)=97.0819$, $p=2.631e-14$; AL: 82.8° vs. 77.2°, $F(2,34)=9.9190$, $p=8.084e-04$; AM: 89.6° vs. 87.5°, $F(2,34)=1.7520$, $p=0.1887$, corrected for multiple-comparisons). PEs thus appear to enhance pattern-separability in lower processing stages by optimizing the available dynamic range, while the top maintains a stable prior. Taken together, our results suggest that pattern separation may be the mechanism underlying the beneficial effects of predictability in IT cortex, decreasing interference between representations of identities. By increasing pattern-separability at lower stages of processing, predictions may thus help in achieving one of the core computations for face/object recognition earlier than in the absence of predictive context.

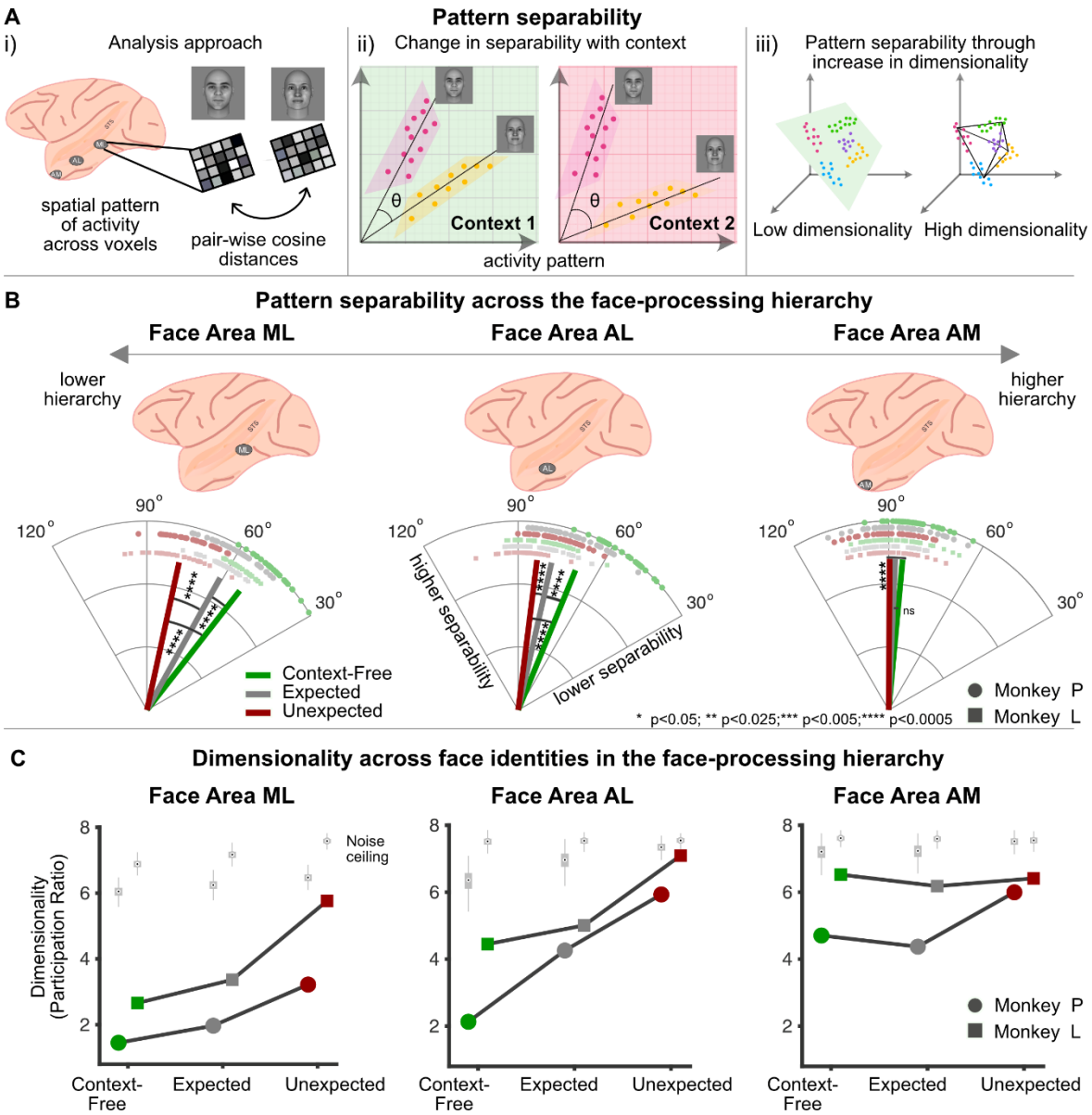

Figure 2: Predictions increase separability in the face-processing hierarchy. ***A.i)*** Pattern-separability was computed as pair-wise cosine distances between activity patterns elicited by all faces presented in each context. Faces were computer generated in Facegen. ***A.ii)*** Neural activity patterns are well separable when there is minimal overlap between population vectors. Separability between the same faces can depend on context. The angle θ between population vectors denotes separability. Faces were computer generated in Facegen. ***A.iii)*** Activity patterns may not be well-separable in low-dimensional spaces. Expanding dimensionality, e.g., by adding feature dimensions, can increase separability. ***B)*** Pattern-separability increases along the face-processing hierarchy from ML (52.1°) to AL (67.1°) to AM (84.7°) in the context-free condition (green). Lower areas in the hierarchy increase pattern-separability when faces are expected (gray; ML 52.1° to 61°; p=1.067e-12; AL: 67.7° to 77.2°, p=2.619e-07), whereas AM shows no difference in separability (84.7° vs. 87.5°, p=0.0644). Prediction violations (red) increase separability in all areas even further (ML: 77.7° vs. 52.1°, p<0.0001; AL: 82.8° vs. 67.7°, p=4.44e-16; AM: 89.6° vs. 84.7°, p=0.0003). Only lower areas have higher separability in the unexpected than in the expected condition (ML: 77.1° vs. 61°, p=2.631e-14; AL: 82.8° vs. 77.2°, p=8.084e-04; AM: 89.6° vs. 87.5°, p=0.1887). Lines depict mean angles, squares and circles pair-wise angles between identities per monkey. ***C)*** Estimated dimensionality in all areas and conditions is lower than that of the noise ceiling (light grey boxplots, all p<9.99e-04). In the context-free condition (green), dimensionality is higher in AM than in lower areas (all p<9.9e-04 for both monkeys). Dimensionality increases when faces are expected

(gray) in ML (Monkey L: p=0.0189, Monkey P: p=0.045) and AL (Monkey L: p=0.004; Monkey P: p<0.0001). Violations (red) expand dimensionality further in ML (Monkey L: p<0.0001, Monkey P: p<0.0001) and AL (Monkey L: p<0.0001; Monkey P: p=0.0001), and less consistently in AM (Monkey P: p<0.0001; Monkey L: p=0.055). Squares and circles depict individual monkeys. All p-values were corrected for multiple comparisons.

How do predictions give rise to increased separability in the hierarchy? We hypothesized that contextual information provided by predictions could add feature dimensions for separability – implying that predictive cues improve separability by increasing dimensionality of neural representational space (Fig. 2Aiii). To evaluate this hypothesis, we quantified dimensionality using the Participation Ratio (PR)[13,14] which describes how evenly the variance of neural activity is spread (i.e., dimensionality is high if variance is spread across all dimensions). PR was computed per face-area and condition across all voxels and stimuli. We first established that PR was interpretable, i.e., well differentiated from noise, in all conditions and face-areas (Fig. 2C; comparison to noise-ceiling for each monkey, all p<9.99e-04, corrected for multiple comparisons). Next, we investigated PR as a function of hierarchy. In the context-free condition, we found higher dimensionality in AM than in lower stages ML and AL (comparison of difference to noise ceiling between areas for each monkey, all p<9.9e-04, corrected for multiple comparisons). For expected faces, dimensionality increased in lower areas AL and ML (that had shown increases in pattern separability) compared to the context-free condition (Monkey L – ML: p=0.0189, AL: p=0.004; Monkey P – ML: p=0.045, AL: p<0.0001, corrected for multiple comparisons). When expectations were violated, dimensionality increased even further compared to the expected condition in ML (Monkey L: p<0.0001; Monkey P: p<0.0001, corrected for multiple comparisons) and AL (Monkey L: p<0.0001; Monkey P: p=0.0001, corrected for multiple comparisons), and slightly less consistently in AM (Monkey P: p<0.0001; Monkey L: p=0.055, corrected for multiple comparisons). Increases in dimensionality and separability across conditions and hierarchy were highly correlated (Spearman's rho=0.95, p=3.5273e-04), suggesting that contextual signals (like predictions and/or PEs) increase pattern-separability through an expansion of representational dimensionality.

Previous studies have shown that deviance between predictions and sensory inputs is signaled as a difference in the magnitude of neural activity between unexpected and expected stimuli[7,16,17]. Indeed, we also found stronger responses in the unexpected than in the expected condition in all three face-areas (Supplementary Fig. S2; all p<3.5e-04). However, this difference in magnitude did not explain the increase in pattern-separability (Spearman's rho=0.0167, p=0.9816) nor dimensionality (rho=0.2667, p=0.4933) with contextual information across the hierarchy. This is in line with previous findings indicating that population magnitude and separability in IT cortex convey different information[18].

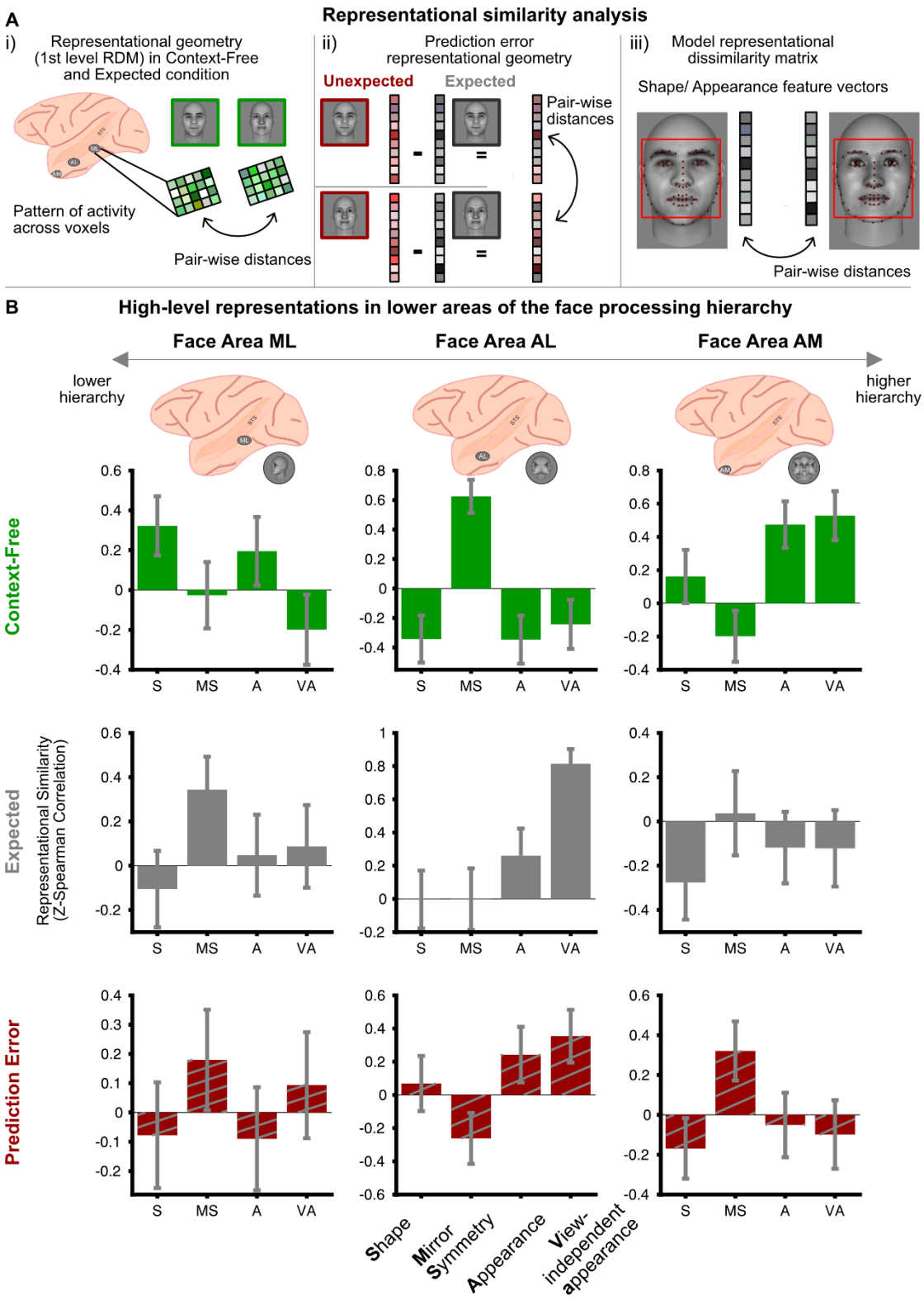

Figure 3: High-level predictions cascade down the face-processing hierarchy. ***A.i.)*** Representational Similarity Analysis (RSA) was performed for each face area1st level representational dissimilarity matrices (RDMs) were computed for the context-free (green) and expected (gray) condition as cosine distances between the stimulus-evoked activity patterns to the different faces shown within each condition. ***A.ii)*** Prediction error (PE) geometry was calculated on the relative activity difference of the same faces in the expected vs. unexpected condition. ***A.iii)*** 1st level RDMs were compared to model RDMs derived from known tuning properties of the face areas (**s**hape, **m**irror-symmetry, **a**ppearance, **v**iew-independent

appearance). Model RDMs were computed on distances between features vectors of the face stimuli. ***B)*** In the context-free condition (green), activity patterns recapitulated known tuning properties of all three face areas (ML: shape vs. 3 others, p=0.0417; AL: mirror-symmetry vs. 3 others, p<0.0001; AM: 2 appearance-based models vs. 2 others, p=0.0019). When faces were expected (gray), lower areas showed increases in higher-order tuning properties compared to the context-free condition (ML: mirror-symmetry, p=0.0460; AL: appearance p=0.0072 and view-independent appearance, p=8.941e-6). This was accompanied by a decrease of feedforward tuning properties in all areas (ML: shape, p=0.0323; AL: mirror-symmetry, p=0.0073; AM: appearance, p=0.0091, view-independent appearance, p=0.0043). Across areas, PE (red) geometry was highly correlated to the geometry in the expected condition (Spearman's rho=0.7273, p=0.01). Compared to the context-free condition, PEs in AL showed increased appearance (p=0.0086) and view-independent appearance tuning (p=0.0079), along with a decrease in mirror-symmetry (p=0.0022). Similarly, AM showed a decrease in appearance (p=0.0178) and view-independent appearance tuning (p=0.0049). Error bars depict the bootstrapped (n=10000) SEM of the Z-transformed 2nd level similarity. See Supplementary Table 1 for individual model fits. All faces were computer generated in Facegen.

How can predictive signals increase separability in lower stages of the hierarchy? One possibility is that lower areas inherit (more separable) representations from higher areas via feedback pathways conveying predictions. Higher areas in the ventral stream contain representations that are not only highly separable, but also invariant, e.g., for view. A consequence of predictions being passed from higher to lower areas is that lower areas should inherit the invariant tuning properties of the areas where the predictions originate. To evaluate this hypothesis, we determined the representational geometry across the face-processing hierarchy using Representational Similarity Analysis (RSA)[19] (Fig. 3Ai). We computed cosine distances between the activity patterns for different face stimuli and compared 1st level RDMs to model RDMs based on known electrophysiological tuning properties of each face area (Fig. 3Aiii, see below). Results were highly correlated across monkeys (Spearman's rho=0.5524; p=2.3830e-04), hence, we report results from averaged 1st level RDMs.

First, in the context-free condition, we established that our approach recapitulated the known electrophysiological tuning properties of face-processing areas[5,6]. As expected, the known tuning properties best explained the representational geometry in all three face areas: view-specific representation of facial shape in ML, mirror-symmetric representation in AL, and view-invariant representation of facial appearance in AM (Fig. 3B; bootstrapped test comparing 2nd level correlation coefficients, one-sided; ML: shape vs. 3 others, difference=0.3153; p=0.0417; AL: mirror symmetry vs. 3 others, difference=0.8364, p<0.0001; AM: 2 appearance-based models vs. 2 others, difference=0.4712, p=0.0019). Next, we tested the hypothesis that under predictable conditions, lower areas exhibit higher-order tuning properties. To this end, we compared representational geometry between the expected and context-free conditions in ML and AL. These lower areas indeed expressed higher-order, abstract representations: ML showed an increase in mirror-symmetric tuning, characteristic of AL (Raghunathan's test, p=0.0460), while AL showed increased tuning for appearance (p=0.0072) and view-independent appearance (p=8.941e-6), characteristics of

AM. Moreover, in the predictable condition we observed a decrease in the classical, feedforward tuning properties in the entire face-processing network, i.e., decreased shape tuning in ML (p=0.0323), decreased mirror-symmetric tuning in AL (p=0.0073), and decreased appearance (p=0.0091) and view-independent appearance tuning (p=0.0043) in AM. Thus, rather than expressing their own feedforward tuning properties, lower areas ML and AL exhibited representations characteristic of higher face-areas when stimuli were predictable. This suggests that high-level predictions cascade-down the entire face-processing hierarchy along feedback pathways, carrying with them a representational format that enables better pattern separation.

Finally, we investigated the representational format of PEs in the unexpected conditions. Predictive processing theory and our results so far suggest that lower areas inherit tuning properties from higher areas, which implies that PEs in lower areas should also express higher-level tuning properties[7]. To test this, we isolated PEs by computing the relative difference between the expected and the unexpected conditions and then applied RSA (Fig. 3Aii). For identity PEs, we found that the pattern of tuning properties was highly correlated with the expected condition (Fig. 3B; Spearman's rho=0.7273, p=0.01), suggesting that the representational geometry evident in the expected condition indeed forms the basis of the computation of PEs when unexpected identities occur. Further mirroring the expected condition, there was an increase in appearance (Raghunathan's test, p=0.0086) and view-independent appearance tuning (p=0.0079) in AL compared to the context-free condition, concomitant to a decrease in mirror-symmetric tuning (p=0.0022). Similarly, appearance (p=0.0178) and view-independent appearance tuning (p=0.0049) again decreased in AM. This suggests that identity PEs in the unexpected identity condition are computed in AL on the basis of appearance tuning originating in AM, i.e., one stage earlier than identity information in the context-free condition. Similarly, and in accordance with our previous electrophysiological results from ML[7], we found significant mirror-symmetric tuning in ML (bootstrapped test, p=0.039) in the unexpected view condition (although this form of tuning did not exhibit a statistically significant increase over the context-free condition; Fisher test, p=0.188). This was accompanied by a decrease of mirror-symmetric tuning in AL (p=0.0003). Overall, this suggests that PEs are computed on the basis of high-level prediction signals in lower areas of the face-processing hierarchy.

Discussion

Taken together, our results reveal that incorporating predictive cues optimizes neural representations for distinguishing facial identities by increasing separability of neural activity

289 patterns already early in the face-processing hierarchy. The improved separability with
290 contextual information comes with an expansion of dimensionality of neural
291 representational space. These effects can be achieved by cascading-down prediction
292 signals via feedback pathways - such that well separated, invariant representations
293 characteristic of higher face-areas are passed down to lower areas. Our work thus proposes
294 a distinct mechanism, i.e., top-down pattern-separability, through which predictions exert
295 their effects.
296 Well separated neural representations are beneficial for processing because they allow for
297 linear readout of information[20,21] and learning new associations[22]. Pattern separation is a
298 well-established computational mechanism, originally attributed to hippocampal and
299 cerebellar circuits[23–25]. Recent proposals assign this computation also to cortex[26] - albeit
300 with different implementations owing to the distinct architectures. We find that predictive
301 information enables high separability in lower areas of the ventral visual stream, which is
302 otherwise only characteristic of the top of this cortical processing hierarchy. Highly
303 separated representations early on in the hierarchy may be beneficial since they facilitate
304 early readout of high-level information like object identity.
305 Theory suggests that highly separable representations can arise from dimensionality
306 expansion[20,21] - providing a larger neural activity space for linear separability. We find that
307 AM at the top of the face-processing hierarchy, which is thought to directly contribute to
308 face individuation, has high dimensionality, similar to the human face-processing
309 hierarchy[15]. This may be related to the inherently high dimensionality of face space required
310 to differentiate individuals[27]. Increases in separability in AL and ML brought about by
311 predictability are accompanied by an expansion of dimensionality. This suggests that the
312 dimensionality of IT cortex activity is in principle flexible and not hardwired. Our findings
313 thus resonate with studies in other species and brain areas which have shown that the
314 dimensionality of neural activity can change on short timescales, e.g., from spontaneous to
315 stimulus-evoked activity[28], or with task engagement[29].
316 Furthermore, we find that in the presence of expectations, lower areas in the face-
317 processing hierarchy express higher-order representations. This suggests that predictions
318 originating from high-level face-processing areas are propagated down to lower areas via
319 feedback connections, in line with predictive processing theories[30]. Predictions lead to a
320 complete transformation of representational geometry, suggesting that they impose the
321 high-level representational format characteristic of the stage of processing at which they
322 are generated, at the expense of the feedforward tuning properties of the respective lower
323 area. This transformation spans the entire face-processing hierarchy: ML inherits tuning
324 properties of AL, and AL inherits tuning properties of AM. This results in the processing of
325 facial appearance, a dimension most closely related to face identification[31], at earlier stages

326 of processing than when no context is available. AM itself no longer shows image-bound
327 properties expressed in the ventral face-processing network, but may inherit yet more
328 abstract, perhaps conceptual representations from the frontal or medial temporal lobe
329 (MTL) with which it is intricately connected[32,33]. This radical transformation between
330 predicted and context-free representations underscores IT's capacity to flexibly and
331 dynamically switch coding regimes on very short timescales, i.e., trial-by-trial, and refutes
332 the notion that tuning properties are hard-wired after years of learning and development.
333 Signaling predictions and deviations from them not only improves sensory encoding but
334 also refines internal models by means of PEs. We find that unexpected stimuli boost
335 pattern-separation in lower face-processing areas beyond the separability found in the
336 expected condition, while not modulating separability at the top of the hierarchy. PEs may
337 refine internal models by increasing separability, in line with memory updating in the MTL[34]
338 and episodic memory enhancement[35,36]. The concomitant increase in dimensionality that
339 we find in the unexpected condition suggests that contextual information provided by
340 predictions introduces additional dimensions into the representational geometry for
341 separability. This would be in line with conjunctive representations between stimulus
342 features and context[18,37], specifically between stimulus features and PEs[7] observed
343 electrophysiologically in IT cortex. Such mixed selectivity could boost pattern-separability
344 and enable more effective readout from high dimensional representational spaces[20].
345 Previous accounts have explained predictions by dampening, i.e., gain modulation[38], or
346 sharpening of highly selective neurons[39]. Dampening predicts amplitude differences
347 between expected and unexpected stimuli; while we find such differences in amplitudes in
348 line with previous studies (Supplementary Fig. S2), they do not explain the differences in
349 separability that we find across the face-processing hierarchy and between conditions,
350 which arise from the multivariate spatial pattern of activity. Sharpening would also predict
351 better discriminability of facial identities as a function of the underlying selectivity; however,
352 sharpening accounts would suggest increased separability solely in AM because only those
353 neurons are highly identity-selective[5]. Furthermore, sharpening may be specific to early
354 visual cortex and/or low dimensional stimuli such as gratings[40,41]. To what extent previously
355 reported sharpening and dampening effects are related to pattern separation remains an
356 important target for future research.
357 Overall, our results suggest a new account of how the ventral stream implements
358 predictions: top-down signals that dynamically transform neural representations into
359 separable and high-dimensional neural geometries. Passing tuning properties down to
360 earlier processing stages allows extraction of high-level, abstract information earlier than
361 what is possible in the absence of predictions, when feedforward computation dominates
362 and signals need to ascend the entire hierarchy to provide information about facial identity.

Predictions thus also free resources for the processing of yet more abstract, less image-bound information in the areas whose feedforward tuning properties are now shifted upstream. Predicting on the basis of invariant tuning properties that abstract over irrelevant features circumvents overfitting to individual images and has an extended temporal horizon compared to pixel-level predictions which are prone to fail in noisy conditions. Our results thus exemplify how feedback connections enable flexible representational spaces that optimally use the computational resources provided by cortical processing hierarchies.

Acknowledgements

We thank W. Freiwald for generously allowing us to collect part of the data in his lab; I. Kagan for support; R. Auksztulewicz for insightful discussions; L. Melloni for comments on the manuscript; A. Gonzalez, L. Burchardt, R. Mielsch, and S. Plewe for help with animal training and care; T. Becker, O. Batura, and A. Schrod for veterinary care; S. Karmacharya, W. Zarco, and R. Ludwig for help with the respective setups; and L. Yin for administrative support. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through an Emmy Noether grant (SCHW1683/2-1) and SFB 1528 - ‘Cognition of Interaction’ (Project-ID 454648639, project A04) to C.M.S.; additional support was provided by an Outgoing Grant from the Leibniz ScienceCampus 'Primate Cognition', Göttingen, Germany to T.N.

Author contributions

T.N.: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Writing - Original Draft; Writing - Review & Editing; Visualization.
C.M.S.: Conceptualization; Methodology; Writing - Original Draft; Writing - Review & Editing; Supervision; Project administration; Funding acquisition.

Declaration of Interest

The funders had no role in study design, data collection and interpretation, decision to publish, or preparation of the manuscript. The authors declare no competing financial or non-financial interests.

Methods

Resource Availability

Lead Contact

Further information and requests for resources should be directed and will be fulfilled by the lead contact, Caspar M. Schwiedrzik (c.schwiedzik@eni-g.de) upon reasonable request.

Material Availability

This study did not generate new or unique reagents.

Data and Code Availability

Data will be made openly available in a repository after publication.

Experimental Model and Study Participant Details

Monkey Subjects and Surgical Procedures

Data for this study were acquired in two adult macaque monkeys (*Macaca mulatta*): one male (Monkey P, 6.5-7 kg, 9 years) and one female (Monkey L, 8.5-9 kg, 10 years). No power calculation was performed to determine the sample size, but was chosen to be in accordance with previous studies[52,53]. We did not have sufficient sample size to investigate differences between female and male macaque monkeys. All procedures with Monkey P were conducted at the German Primate Center, Göttingen and approved by the responsible regional government office (Niedersächsisches Landesamt für Verbraucherschutz und Lebensmittelsicherheit - LAVES). The animal was either housed in pairs or groups, adhering to German and European regulations. The facility provided a stimulating environment for the animals, including various toys and structures, natural and artificial lighting, and access to outdoor spaces, surpassing the space requirements outlined in European regulations. All procedures with Monkey L complied with 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 21104-H USDA) and Weill Cornell Medicine (protocol number 2010-0029, where

magnetic resonance imaging (MRI) was performed. The animal was housed with other macaque monkeys and had access to a stimulating environment including various toys for enrichment. The well-being of both the animals, both psychologically and medically, was closely monitored on a daily basis by veterinarians, animal facility personnel, and scientists. Both animals were subjected to surgical procedures where MRI-compatible cranial head-posts were implanted, embedded within bone cement, anchored with MR-compatible ceramic screws under general anesthesia and sterile conditions[7]. The animals underwent extensive training using positive reinforcement[54] to become accustomed to entering and remaining seated in a horizontal primate chair. Their head position was stabilized using the implanted head-posts, allowing for cleaning of the implants, accurate recording of gaze/pupil diameter, and minimizing motion artifacts during MRI experiments.

Method Details

MRI Data Acquisition

MRI data were acquired on 3T scanners (Monkey P: MAGNETOM-Prisma; Monkey L: MAGNETOM-Prisma_fit, Siemens Healthineers, Erlangen, Germany). Anatomical images were obtained for each animal in a separate session using a T1-weighted magnetization-prepared rapid gradient echo (MPRAGE) sequence (Monkey P: field of view [FOV] 128 mm, voxel size = 0.5 × 0.5 × 0.5 mm, repetition time [TR] = 2.7 s, echo time [TE] = 2.96 ms, inversion time [TI] = 850 ms, bandwidth [BW] = 220 Hz/Px, flip angle [FA] = 8 degrees, 240 slices, 11 cm loop coil, Siemens Healthineers; Monkey L: FOV = 128 mm, voxel size = 0.5 × 0.5 × 0.5 mm, TR = 2.7 s, TE = 2.99 ms, TI = 868 ms, BW = 230 Hz/Px, FA = 9 degrees, 240 slices, custom 1-channel receive coil L. Wald, MGH Martinos Center for Biomedical Imaging) while the monkeys were anesthetized (isoflurane 1.5%–2%) and positioned in an MR-compatible stereotactic frame (Kopf Instruments). Functional images were acquired using custom 4 or 8 channel phased-array receive surface coils with a horizontally oriented single loop transmit coil (H. Kolster, Windmiller Kolster Scientific, and L. Wald, MGH Martinos Center for Biomedical Imaging) while the monkeys were in a sphinx position. Each functional time series consisted of whole-brain gradient-echo planar images (EPI; FOV = 96 mm, voxel size = 1.2 × 1.2 × 1.2 mm, TR = 2 s, TE = 27 ms, BW = 1302 Hz/Px, echo spacing [ESP]= 0.93 ms, FA = 76 degrees, 43 slices) acquired in interleaved order with two times generalized autocalibrating partially parallel acquisitions (GRAPPA) acceleration. All functional images in the Statistical Learning Paradigm (see below) for both monkeys and Face Localizer (see below) in Monkey P were acquired with the above sequence parameters. Functional images for Face Localizer in Monkey L were acquired with the

contrast agent ferumoxytol (Molday ION, BioPAL, Worcester, USA), and hence slightly different sequence parameters (FOV = 96 mm, voxel size = 1.2 × 1.2 × 1.2 mm, TR = 2.25 s, TE = 17 ms, BW = 1525 Hz/Px, FA = 79 degrees, 45 slices). Right before every scanning session, ferumoxytol was injected into the saphenous vein. The amount administered ranged from 9 mg of Fe per kg of the animal's body weight during the first scan to 6 mg on following days to account for the functional half-life of the contrast agent. On every session, field maps (FOV = 96 mm, voxel size = 1.2 × 1.2 × 2.4 mm, TR = 0.7 s, TE1 = 5 ms, TE2 = 7.46 ms, FA = 60 degrees, 22 slices) were also acquired which were later used for EPI undistortion[55].

Face Localizer

We use a standard face localizer to localize the face patches ML, AL, and AM[56]. Subjects fixated on a central white dot while we displayed images of human/monkey faces, body parts/headless bodies, man-made objects, and fruits, alternating with baseline periods in a block design (FOB). Each block lasted 24-30 s. Fluid reward was given after variable 2-4 s periods, contingent on gaze staying within 2 degrees of the fixation dot. Analysis included runs where subjects achieved ≥80-85% fixation stability. Visual stimulation and reward were controlled using Psychtoolbox[48] for Monkey P and Presentation (v19, Neurobehavioral Systems) for Monkey L. Stimuli were projected on a back-projection screen using a video projector (Monkey P: Epson EB-G5600, refresh rate 60 Hz, resolution 1920 × 1080 px; Monkey L: NEC NP3250, refresh rate 60 Hz, resolution 1024 × 768 px) with a custom lens. Eye position was measured using a video-based eye tracking system (Monkey P: SR Research Eyelink 1000 NHP Long Range Optics at 1000 Hz, Monkey L: ISCAN at 120 Hz).

Statistical Learning Paradigm

For the statistical learning experiment, we generated 30 three-dimensional faces with neutral expression and no hair using FaceGen Modeller Pro 3.27 (Singular Inversions). Seven views (0°, 30°, 45°, 60°, 300°, 315°, 330°) were extracted for each face. The faces were counterbalanced for their gender and skin texture was added such that the right profile image was not a simple mirror-version of the left-profile image. The images were converted to grayscale and then luminance normalized using the SHINE toolbox[49]. The training set for the statistical learning experiment consisted of 18 facial identities with 6 images of frontal, right and left profile-view each. These were split into 9 predictor-successor pairs such that the head orientations were balanced within predictors and successors, respectively[7].
Prior to the training phase, monkeys were familiarized with the entire set of 30 faces in all the 7 views in randomized order such that no image was novel to the monkey. The monkeys

were then exposed to the face-pairs to establish associations between the predictor and the successor images. Pairs were arranged such that one identity-view combination would uniquely predict one other identity-view combination (Supplementary Fig. S3). Training was conducted with sequentially presented face-pairs for at least 3 weeks for both monkeys. The sequence of pairs was designed such that the transitional probability within a pair was 100% and the transitional probability between pairs was kept at a minimum.

In the early phase of training, each image was presented foveally (8 dva in size) for 500 ms with an inter-stimulus interval of 500 ms. Longer baselines (up to 9.5 s) and spatial jittering (~1 dva) from the foveal position were gradually introduced over the training period. This was done to accustom the animals to the task design to be used during the fMRI experiments. The monkey was rewarded with water/juice every 2.5-3.8 s for maintaining fixation within a 2-3 dva window around the fixation dot presented at the center of the screen. To control visual stimulation and reward delivery, we used mWorks (https://mworks.github.io/) and a custom-built input-output device (R. Brockhausen) using Teensy 3.6 (PJRC: Electronic Projects) for Monkey P, and Presentation (v20.2, Neurobehavioral Systems) with a commercially available data acquisition device (Measurement Computing USB 1208FS) for Monkey L. Stimuli were presented using a video projector for Monkey P (Barco F22 WUXGA; refresh rate, 60 Hz; resolution, 1920 × 1080 px) and displayed on a monitor for Monkey L (Dell E2011H, refresh rate 60 Hz, resolution 1920 × 1080 px). Eye position was measured using video-based eye tracking (Monkey P: SR Research Eyelink 1000 at 1000 Hz, Monkey L: ISCAN at 120 Hz).

The testing phase of the statistical learning paradigm took place while fMRI data were acquired (Monkey L: 81 runs; Monkey P: 56 runs). In this phase, learned face-pairs were presented such that one face predicted the next; violations of the predictions (40% of all trials) were introduced in the second image of a face-pair. The violations entailed a successor image with a different identity (but with the expected view) or a different view (but with the expected identity) than during the training phase. To create the identity violation conditions, we systematically recombined the trained predictor image of a pair with that of another successor image to generate unexpected pairs. To create the view violation conditions, we presented the predicted identity but with a different head orientation than during training. The experimenter was not blinded to the experimental conditions. A fixation dot was continuously presented at the center of the screen and the monkey was rewarded for maintaining fixation within a 2-3 dva fixation window every 2.2-3.6 s. Reward was thus not contingent upon image presentation but entirely on fixation performance. We used a rapid-event related design: each face image was presented for 500 ms, aligned to the onset of volume acquisition. To facilitate single-trial deconvolution of the blood oxygenation level

525 dependent (BOLD) response, baseline durations were jittered between 1.5, 3.5 and 5.5 s
526 for intra-pair and 5.5, 7.5 and 9.5 s for inter-pair intervals. The face images were spatially
527 jittered within ~1 dva around the fixation dot to avoid low-level adaptation effects.

528 Quantification and Statistical Analyses

529 MRI processing

530 Anatomical images were intensity normalized, skull-stripped, and segmented using
531 Freesurfer (v5.3, https://surfer.nmr.mgh.harvard.edu/)[47]. Preprocessing of fMRI data was
532 done using Freesurfer's functional analysis stream, FS FAST, AFNI
533 (https://afni.nimh.nih.gov)[42] and the JIP toolkit (https://www.nitrc.org/projects/jip). Slice-
534 wise motion-correction was done within each run using AFNI's 3dAllineate with terms for
535 cubic warping in the phase encoding direction, as well as shifts, rotations, scaling, and
536 skewing, followed by slice-time correction using FS FAST. Geometric distortions of the
537 functional volumes were corrected by means of the field maps, followed by mutual
538 information-based non-linear alignment to the high-resolution anatomical scans as
539 implemented in JIP.
540 Statistical modeling was done using a General Linear Model (GLM) in FS FAST. Predictors
541 of interest were convolved with the canonical hemodynamic response function (HRF) for
542 BOLD for all experiments except for Monkey L's Face Localizer data where HRF
543 parameters were appropriate for ferumuxytol[57]. Motion parameters, fixation stability, and
544 reward were included as nuisance regressors. The first five volumes of each functional run
545 were excluded to prevent T1 saturation artifacts, and detrending (up to 2nd degree
546 polynomials) was applied within the GLM. To identify the face-areas from the FOB data, we
547 contrasted faces vs. all other visual categories[58]. Regions of interest (ROIs) in each
548 hemisphere were defined as the intersection of a sphere around the peak voxel identified
549 by this contrast and the gray matter, resulting in ROIs of ~100 voxels. Since face-processing
550 is not lateralized in macaque monkeys[59], we combined the voxels in each of the face-areas
551 from the two hemispheres for all our analyses (Monkey P: ML 204, AL 185, AM 187; Monkey
552 L: ML 198, AL 190, AM 204 voxels).
553 For the testing phase of the statistical learning experiment, we resolved trial-specific
554 activation using the least squares-single approach[60]. In this approach, a separate GLM is
555 calculated for each trial and the design matrix contains two main task-related regressors:
556 the trial-of-interest was defined as the first regressor and all other trials in a run, irrespective
557 of the experimental conditions, were simultaneously defined as the second (nuisance)
558 regressor. This was done iteratively for every stimulus within a run. The regression

559 coefficient for the trial of interest was contrasted with the baseline and the t-statistic
560 obtained for each trial was used for further analyses[61].

561

562 Pattern separability

563 Pattern-separability was computed across facial identities using cosine distances, a
564 measure commonly used in the pattern separation literature[62,63]. Cosine distances indicate
565 the vector direction of a population response irrespective of amplitude or sample size
566 differences in the experimental conditions[64,65] and have thus been argued to be particularly
567 sensitive to the tuning of neural populations in the context of fMRI[66]. To compute cosine
568 distances, we extracted t-statistics from all trials and conducted multivariate noise-
569 normalization using the covariance matrix of the residuals from the GLM done on every
570 trial[61]. Cosine distances were then converted to angles by computing the inverse cosine.
571 This was done for each stimulus in all ROIs and conditions, respectively. All nine facial
572 identities that were presented as the first stimulus in a pair were used to estimate the
573 separability in the 'Context-Free' condition. The first image in a pair cannot be anticipated
574 on the basis of the preceding image and is thus processed in the absence of contextual
575 predictions. All nine trained successor stimuli were used for the 'Expected' condition. The
576 distances between the same nine stimuli when presented as an expectation violation were
577 used to quantify the separability in the 'Unexpected' condition. Due to the circular nature of
578 angles, circular statistics[44] were used to compare conditions and ROIs, respectively. We
579 used the Hotelling paired sample test for equal angular means[67] for paired comparisons,
580 and the William-Watson test for independent samples comparisons[68]. All p-values were
581 corrected for multiple comparisons using the Bonferroni-Holm method[69]. All the above
582 analyses were done using The Decoding Toolbox[50] in MATLAB (The MathWorks, Inc.,
583 R2018B) and custom MATLAB scripts.

584 Dimensionality

585 To estimate dimensionality of neural representational space in a given ROI in the different
586 experimental conditions, we computed the Participation Ratio (PR)[13,14,45]. PR is the ratio of
587 the square of the first moment and the second moment of the eigenvalue probability density
588 function and quantifies how evenly variance of activity is spread across stimuli; i.e., if
589 variance is widely spread, PR is high. We estimated PR using whitened t-statistics across
590 voxels and stimuli in each condition and ROI. Hence, our measure of dimensionality is
591 constrained by the number of stimuli. To determine whether the dimensionality estimates
592 are interpretable, we compared the estimated dimensionality to that of noise. Noise ceilings

were estimated by synchronized permutation (1000) of the stimuli and voxels for each condition and ROI. We then tested whether the estimated dimensionality is lower than the 95% confidence interval (CI) of the noise distribution. Similarly, pairwise differences between the dimensionality across ROIs or conditions were assessed by comparing whether the increase in estimated dimensionality was higher than the 95% CI of the noise distribution. All p-values were corrected for multiple comparisons using the Bonferroni-Holm method[69]. PR was quantified using publicly available[45] Python scripts and statistically analyzed using custom MATLAB scripts.

Population response magnitude

Population response magnitudes[18] were quantified as the Euclidean (L2) norm across all voxels in a given ROI for each stimulus separately and then averaged. This served to disentangle differences in the magnitude of the IT population response from angular differences in stimulus identity as it has been shown that population response magnitude and population response vector direction can contain differential information[70]. We computed the Euclidean (L2) norm from the t-statistics per stimulus without further normalization so as not to discard any relevant magnitude differences, as recommended in[18]. Paired t-tests were used to assess differences in population response magnitude across expected and unexpected conditions, and corrected for multiple comparisons using the Bonferroni-Holm method[69].

Representational Similarity Analysis

Representational Similarity Analysis (RSA)[19] was conducted to assess tuning properties of face-areas and how they change with experimental conditions. We first determined whether we could replicate the well-known tuning properties of the face-areas using BOLD signals, relative to the electrophysiological "gold standard"[5,6]. To this end, we computed 1st level representational dissimilarity matrices (RDMs) using cosine distances (as described in the separability section above) on the pattern of BOLD responses across voxels in each of the ROIs. We then compared these 1st level RDMs to representations predicted by previous electrophysiology studies (Model/Hypothetical RDMs).

The choice of model RDMs was based on previous studies that show that ML has a view-specific representation of facial shape, AL has mirror-symmetric representation, and AM has view-invariant representation of the facial appearance[5,6]. To compute hypothetical RDMs for shape and appearance, the Active Appearance Model[51,71] (https://www.menpo.org/) was trained using 871 pre-annotated images and used to fit each

627 of our face stimuli to obtain 68 shape landmarks[72]. We then computed Spearman rank
628 correlation distances between the feature vector of landmark positions for all the stimuli to
629 compute a hypothetical shape RDM. We used 45° and 135° faces to extract all 68 shape
630 landmarks to ascertain a full set of shape landmarks for all the facial views such that
631 similarity between the faces could be computed considering all views. Next, to compute an
632 appearance representational space, each face was smoothly warped to the average shape
633 template made from the training set, using spline interpolation. This warped image was then
634 normalized for the mean and reshaped to obtain an appearance feature vector. We then
635 computed Spearman rank correlation distances between the feature vector of the
636 appearance feature vector for all the stimuli to obtain a hypothetical appearance RDM. Note
637 that this appearance model RDM contains a component of facial view and is different from
638 the shape-free appearance model referred to in[6] where the authors computed appearance
639 representations on each view separately. Therefore, we additionally computed a model
640 RDM using only the frontal-view versions of the same nine facial identities and performed
641 the same procedure as above to obtain the hypothetical appearance RDM. Although this
642 hypothetical RDM is made with only frontal faces, correlating it with an empirical RDM
643 based on three different views is what makes this a view-independent/invariant appearance
644 model. Lastly, we designed a hypothetical mirror-symmetry RDM by assigning the distance
645 within each view and between left and right profile faces to be zero[5].
646 Comparisons of model RDMs with empirical, 1st level RDMs were carried out using
647 Spearman rank correlation[19], partialling out low-level similarities[73] using a Gabor wavelet
648 pyramid as a model of early visual cortex[74,75]
649 (https://github.com/daseibert/image_similarity_toolbox). To test whether the known
650 electrophysiological properties of each face-area best explain the representations in each
651 of the three face-areas, we compared the model fit of the candidate RDM (e.g., Fisher *z*-
652 transformed Spearman rank correlation between Mirror-symmetry model RDM and the 1st
653 level RDM in AL) to the average model fits of the three other models (e.g., shape,
654 appearance, and view-independent appearance). This was done using bootstrapping
655 (10000 samples, one-sided) in the context-free condition when no predictability exists, and
656 which corresponds most closely to the electrophysiological studies that established the
657 tuning properties of the face-areas. To test whether tuning properties changed as a function
658 of experimental condition, i.e., whether there is an increase in higher-order tuning properties
659 in lower areas and a decrease in feedforward tuning properties in higher areas, respectively,
660 we compared the model fit of the candidate RDM in the condition with no context to that of
661 its model fit in the expected/unexpected identity conditions. This was done by using
662 Raghunathan's test[76] (one-sided) for dependent, non-overlapping correlation coefficients
663 which also takes intercorrelations between the model and experimental RDMs into account.

To study the representational geometry of PEs, we first computed the relative difference between the expected and the respective unexpected condition (view, identity). This effectively isolates PEs, which are thought to signal the difference between expected and unexpected stimuli with increases in neural activity[7]. Following this, we computed pairwise cosine distances between the change in activity pattern and determined 2nd level similarity by means of RSA. For the unexpected identity condition, cosine distances were computed across the different identities before the 2nd level RSA was conducted. For the unexpected view condition, we calculated the difference between the expected identity and the view violations of the corresponding identity (both unexpected views of a given identity separately). Then, the 1st level RDM was computed across the different face images, followed by 2nd level RSA (the Gabor wavelet pyramid was not partialled out for this analysis). Since PEs are thought to be computed on the representational format of expectations, we hypothesized that PEs also lead to an increase in high-level properties in lower areas of the hierarchy. We assessed increases in high-level tuning properties and decrease in feedforward tuning properties between unexpected identity and context-free conditions using Raghunathan's test for dependent, non-overlapping correlations[76]. To compare the context-free with the unexpected view condition, we used Fisher's test for independent, non-overlapping correlations[77]. These analyses were carried out in R (v3.6.2) using the 'cocor' toolbox[43].

Pupil entrainment

Pupil entrainment was used to determine whether the face-pair structure was learned, in accordance with previous studies establishing neural[78,79] and pupillary entrainment[9] as a signature of statistical learning. During training, face images were presented at a rate of 1 Hz (the image frequency), resulting in a pair frequency of 0.5 Hz. To determine pupil entrainment to face pairs, we computed inter-trial phase coherence[80] (ITC). Only runs in which fixation stability was >85 % for Monkey P and >80% for Monkey L were selected. Preprocessing of pupil area data was done in accordance with previously published work on pupil entrainment to face-pairs[9]: blinks and missing data were linearly interpolated; outliers in pupil area were detected using median absolute deviation (cutoff 3.5 for Monkey P and 1.5 for Monkey L) and linearly interpolated. This was followed by low-pass filtering at 5 Hz using a onepass-zerophase Kaiser-windowed sinc finite impulse response (FIR) filter (filter order 1812, transition width 2.0 Hz, pass band 0-4.0 Hz, stop band 6.0-500 Hz, maximal pass band deviation 0.0010 (0.10%), stop band attenuation -60 dB)[81]. Subsequently, data were detrended per run and per block; the average pupil area over the baseline of 2 s before the beginning of stimulation in each block was subtracted. Data from

Monkey P were then downsampled to 500 Hz and Monkey L was retained at the acquisition frequency (120 Hz). Pseudo-trials of 32 seconds were made from continuous data, time-locked to the first stimulus in a face-pair. The pseudo trials had 30 second overlap for Monkey P and 26 second overlap for Monkey L. An additional baseline correction was performed per pseudo-trial by subtracting an average of the pupil area over the 2 s pre-block baseline. Data for Monkey L which was acquired at a low sampling rate was additionally detrended per pseudo-trial. We then computed a trial-by-trial Fourier transform using discrete prolate spheroidal sequences (DPSS) as tapers with a spectral resolution of 0.0625 Hz, using the Matlab toolbox Fieldtrip (v20190329; http://www.fieldtriptoolbox.org/)[46]. The resulting complex spectra were used to calculate ITC. ITC at the pair-frequency was normalized by dividing the mean ITC of 4 surrounding frequency bins (2 above, 2 below). Normalized ITC values above 1 thus indicate entrainment at the respective frequency. Statistical significance of ITC at the pair-frequency was determined using a one-sided t-test against 1 in both monkeys. In addition, to assess whether ITC at the pair frequency increased from early to late phases of training, we compared the first and second half of training sessions, respectively, in monkey P (unpaired, one-sided t-test). Due to technical issues and time constraints, pupil data for monkey L was only available from the late training phase.

References

1. H. Lamba, A. Sarkar, M. Vatsa, R. Singh, and A. Noore (2011). Face recognition for look-alikes: A preliminary study. In 2011 International Joint Conference on Biometrics (IJCB), pp. 1–6. 10.1109/IJCB.2011.6117520.
2. Summerfield, C., and de Lange, F.P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nat. Rev. Neurosci. *15*, 745–756. 10.1038/nrn3838.
3. DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn. Sci. *11*, 333–341. 10.1016/j.tics.2007.06.010.
4. Spratling, M.W. (2017). A review of predictive coding algorithms. Brain Cogn. *112*, 92–97. 10.1016/j.bandc.2015.11.003.
5. Freiwald, W.A., and Tsao, D.Y. (2010). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. Science *330*, 845–851. 10.1126/science.1194908.
6. Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028. 10.1016/j.cell.2017.05.011
7. Schwiedrzik, C.M., and Freiwald, W.A. (2017). High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. Neuron *96*, 89-97.e4. 10.1016/j.neuron.2017.09.007.
8. Meyer, T., and Olson, C.R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. Proc. Natl. Acad. Sci. *108*, 19401–19406. 10.1073/pnas.1112895108
9. Schwiedrzik, C.M. and Sudmann, S.S. (2020). Pupil diameter tracks statistical structure in the environment to increase visual sensitivity. J. Neurosci. 40, 4565. 10.1523/JNEUROSCI.0216-20.2020.
10. Rust, N. C. and DiCarlo, J.J. (2010). Selectivity and Tolerance (“Invariance”) Both Increase as Visual Information Propagates from Cortical Area V4 to IT. J. Neurosci. *30*, 12978. 10.1523/JNEUROSCI.0179-10.2010.
11. Huang, Y., and Rao, R.P.N. (2011). Predictive coding. WIREs Cogn. Sci. *2*, 580–593. 10.1002/wcs.142.
12. Srinivasan, M.V., Laughlin, S.B., and Dubs, A. (1982). Predictive Coding: A Fresh View of Inhibition in the Retina. Proc. R. Soc. Lond. B Biol. Sci. *216*, 427–459.
13. Gao, P., Trautmann, E., Yu,B., Santhanam,G., Ryu,S., Shenoy,K., and Ganguli,S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv, 214262. 10.1101/214262.

14. Recanatesi, S., Bradde, S., Balasubramanian, V., Steinmetz, N.A., and Shea-Brown, E. (2022). A scale-dependent measure of system dimensionality. Patterns *3*, 100555. 10.1016/j.patter.2022.100555.
15. Sheng, J., Zhang, L., Liu, C., Liu, J., Feng, J., Zhou, Y., Hu, H., and Xue, G. Higher-dimensional neural representations predict better episodic memory. Sci. Adv. *8*, eabm3829. 10.1126/sciadv.abm3829.
16. Alink, A., Schwiedrzik, C.M., Kohler, A., Singer, W., and Muckli, L. (2010). Stimulus Predictability Reduces Responses in Primary Visual Cortex. J. Neurosci. *30*, 2960. 10.1523/JNEUROSCI.3730-10.2010.
17. Kaposvari, P., Kumar, S., and Vogels, R. (2018). Statistical Learning Signals in Macaque Inferior Temporal Cortex. Cereb. Cortex *28*, 250–266. 10.1093/cercor/bhw374.
18. Jaegle, A., Mehrpour, V., Mohsenzadeh, Y., Meyer, T., Oliva, A., and Rust, N. (2019). Population response magnitude variation in inferotemporal cortex predicts image memorability. eLife *8*, e47596. 10.7554/eLife.47596.
19. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4. 10.3389/neuro.06.004.2008.
20. Fusi, S., Miller, E.K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Neurobiol. Cogn. Behav. *37*, 66–74. 10.1016/j.conb.2016.01.010.
21. Cayco-Gajic, N.A., and Silver, R.A. (2019). Re-evaluating Circuit Mechanisms Underlying Pattern Separation. Neuron *101*, 584–602. 10.1016/j.neuron.2019.01.044.
22. Barlow, H., and Foldiak, P. (1989). Adaptation and decorrelation in the cortex. In The Computing Neuron, C. Miall, R. Durbin, and G. Mitchison, eds. (Addison Wesley), pp. 54–72.
23. Marr, D. (1969). A theory of cerebellar cortex. J. Physiol. *202*, 437–470. 10.1113/jphysiol.1969.sp008820.
24. Marr, D., and Brindley, G.S. (1997). Simple memory: a theory for archicortex. Philos. Trans. R. Soc. Lond. B Biol. Sci. *262*, 23–81. 10.1098/rstb.1971.0078.
25. Albus, J.S. (1971). A theory of cerebellar function. Math. Biosci. *10*, 25–61. 10.1016/0025-5564(71)90051-4.
26. Amer, T., and Davachi, L. (2023). Extra-hippocampal contributions to pattern separation. eLife *12*, e82250. 10.7554/eLife.82250.
27. M. Meytlis, and L. Sirovich (2007). On the Dimensionality of Face Space. IEEE Trans. Pattern Anal. Mach. Intell. *29*, 1262–1267. 10.1109/TPAMI.2007.1033.
28. Mazzucato, L., Fontanini, A., and La Camera, G. (2016). Stimuli Reduce the Dimensionality of Cortical Activity. Front. Syst. Neurosci. *10*. 10.3389/fnsys.2016.00011
29. Yao, J.D., Zemlianova, K.O., Hocker, D.L., Savin, C., Constantinople, C.M., Chung, S.,

789 and Sanes, D.H. (2023). Transformation of acoustic information to sensory decision
790 variables in the parietal cortex. Proc. Natl. Acad. Sci. *120*, e2212120120.
791 30. Rao, R.P.N., and Ballard, D.H. (1999). Predictive coding in the visual cortex: a functional
792 interpretation of some extra-classical receptive-field effects. Nat. Neurosci. *2*, 79–87.
793 10.1038/4580.
794 31. Andrews, T.J., Rogers, D., Mileva, M., Watson, D.M., Wang, A., and Burton, A.M.
795 (2023). A narrow band of image dimensions is critical for face recognition. Vision Res.
796 *212*, 108297. 10.1016/j.visres.2023.108297.
797 32. Schwiedrzik, C.M., Zarco, W., Everling, S., and Freiwald, W.A. (2015). Face patch
798 resting state networks link face processing to social cognition. PLoS Biol. *13*, e1002245.
799 10.1371/journal.pbio.1002245
800 33. Grimaldi, P., Saleem, K.S., and Tsao, D. (2016). Anatomical Connections of the
801 Functionally Defined “Face Patches” in the Macaque Monkey. Neuron *90*, 1325–1342.
802 10.1016/j.neuron.2016.05.009.
803 34. Bein, O., Duncan, K., and Davachi, L. (2020). Mnemonic prediction errors bias
804 hippocampal states. Nat. Commun. *11*, 3451. 10.1038/s41467-020-17287-1.
805 35. Frank,D., Montemurro,M.A., and Montaldi, D. (2020). Pattern Separation Underpins
806 Expectation-Modulated Memory. J. Neurosci. *40*, 3455. 10.1523/JNEUROSCI.2047-
807 19.2020.
808 36. Barron, H.C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: A
809 predictive coding account. Prog. Neurobiol. *192*, 101821.
810 10.1016/j.pneurobio.2020.101821.
811 37. She,L., Benna,M.K., Shi,Y., Fusi,S., and Tsao, D.Y. (2021). The neural code for face
812 memory. bioRxiv, 2021.03.12.435023. 10.1101/2021.03.12.435023.
813 38. Kumar, S., Kaposvari, P., and Vogels, R. (2017). Encoding of Predictable and
814 Unpredictable Stimuli by Inferior Temporal Cortical Neurons. J. Cogn. Neurosci. *29*,
815 1445–1454. 10.1162/jocn_a_01135.
816 39. Kok, P., Jehee, J.F.M., and de Lange, F.P. (2012). Less Is More: Expectation Sharpens
817 Representations in the Primary Visual Cortex. Neuron *75*, 265–270.
818 10.1016/j.neuron.2012.04.034.
819 40. Richter, D., Ekman, M. and de Lange, F.P. (2018). Suppressed Sensory Response to
820 Predictable Object Stimuli throughout the Ventral Visual Stream. J. Neurosci. *38*, 7452.
821 10.1523/JNEUROSCI.3421-17.2018.
822 41. Richter, D., Heilbron, M., and de Lange, F.P. (2022). Dampened sensory
823 representations for expected input across the ventral visual stream. Oxf. Open Neurosci.
824 *1*, kvac013. 10.1093/oons/kvac013.
825 42. Cox, R.W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic

Resonance Neuroimages. Comput. Biomed. Res. *29*, 162–173. 10.1006/cbmr.1996.0014.

43. Diedenhofen, B., and Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. PloS One *10*, e0121945. 10.1371/journal.pone.0121945

44. Berens, P. (2009). CircStat: A MATLAB Toolbox for Circular Statistics. J. Stat. Softw. *31*, 1–21. 10.18637/jss.v031.i10.

45. Stephenson, C., Feather, J., Padhy, S., Elibol, O., Tang, H., McDermott, J., and Chung, S. (2019). Untangling in invariant speech recognition. Adv. Neural Inf. Process. Syst. *32*. arXiv:2003.01787v1

46. Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2010). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput. Intell. Neurosci. *2011*, 156869. 10.1155/2011/156869.

47. Fischl, B. (2012). FreeSurfer. 20 YEARS FMRI *62*, 774–781. 10.1016/j.neuroimage.2012.01.021.

48. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436. 10.1163/156856897X00357.

49. 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. 10.3758/BRM.42.3.671.

50. Hebart, M.N., Görgen, K., and Haynes, J.-D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Front. Neuroinformatics *8*. 10.3389/fninf.2014.00088

51. Alabort-i-Medina, J., Antonakos, E., Booth, J., Snape, P., and Zafeiriou, S. (2016). The Menpo Project. SIGMultimedia Rec *8*, 4–9. 10.1145/2982857.2982858.

52. Klink, P.C., Chen, X., Vanduffel, W., and Roelfsema, P.R. (2021). Population receptive fields in nonhuman primates from whole-brain fMRI and large-scale neurophysiology in visual cortex. eLife *10*, e67304. 10.7554/eLife.67304.

53. Cui, D., Sypré, L., Vissers, M., Sharma, S., Vogels, R., and Nelissen, K. (2023). Categorization learning induced changes in action representations in the macaque STS. NeuroImage *265*, 119780. 10.1016/j.neuroimage.2022.119780.

54. Prescott, M.J., and Buchanan-Smith, H.M. (2003). Training Nonhuman Primates Using Positive Reinforcement Techniques. J. Appl. Anim. Welf. Sci. *6*, 157–161. 10.1207/S15327604JAWS0603_01.

55. Jezzard, P., and Balaban, R.S. (1995). Correction for geometric distortion in echo planar images from B0 field variations. Magn. Reson. Med. *34*, 65–73. 10.1002/mrm.1910340111.

56. Fisher, C., and Freiwald, W.A. (2015). Contrasting Specializations for Facial Motion within the Macaque Face-Processing System. Curr. Biol. *25*, 261–266. 10.1016/j.cub.2014.11.038.

57. Leite, F.P., Tsao, D., Vanduffel, W., Fize, D., Sasaki, Y., Wald, L.L., Dale, A.M., Kwong, K.K., Orban, G.A., Rosen, B.R., et al. (2002). Repeated fMRI Using Iron Oxide Contrast Agent in Awake, Behaving Macaques at 3 Tesla. NeuroImage *16*, 283–294. 10.1006/nimg.2002.1110.

58. 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. 10.1126/science.1157436.

59. Tsao, D.Y., Moeller, S., and Freiwald, W.A. (2008). Comparing face patch systems in macaques and humans. Proc. Natl. Acad. Sci. *105*, 19514–19519. 10.1073/pnas.0809662105.

60. Mumford, J.A., Turner, B.O., Ashby, F.G., and Poldrack, R.A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. NeuroImage *59*, 2636–2643. 10.1016/j.neuroimage.2011.08.076.

61. Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. NeuroImage *137*, 188–200. 10.1016/j.neuroimage.2015.12.012.

62. Madar, A.D., Ewell, L.A., and Jones, M.V. (2019). Pattern separation of spiketrains in hippocampal neurons. Sci. Rep. *9*, 5282. 10.1038/s41598-019-41503-8.

63. Bird, A.D., Cuntz, H., and Jedlicka, P. (2023). Robust and consistent measures of pattern separation based on information theory and demonstrated in the dentate gyrus. bioRxiv, 2022.11.04.515143. 10.1101/2022.11.04.515143.

64. Chou, E.P., and Hsu, S.-M. (2018). Cosine similarity as a sample size-free measure to quantify phase clustering within a single neurophysiological signal. J. Neurosci. Methods *295*, 111–120. 10.1016/j.jneumeth.2017.12.007.

65. Diedrichsen, J., Berlot, E., Mur, M., Schütt, H., Shahbazi, M., and Kriegeskorte, N. (2021). Comparing representational geometries using whitened unbiased-distance-matrix similarity. Neurons Behav. Data Anal. Theory *5*. 10.51628/001c.27664.

66. Ramírez, F.M. (2018). Orientation Encoding and Viewpoint Invariance in Face Recognition: Inferring Neural Properties from Large-Scale Signals. The Neuroscientist *24*, 582–608. 10.1177/1073858418769554.

67. Zar, J.H. (1999). Biostatistical Analysis (Prentice Hall).

68. Watson, G.S., And Williams, E.J. (1956). On The Construction Of Significance Tests On The Circle And The Sphere. Biometrika *43*, 344–352. 10.1093/Biomet/43.3-4.344.

69. Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scand. J.

Stat. *6*, 65–70.
70. Rust, N.C., and Cohen, M.R. (2022). Priority coding in the visual system. Nat. Rev. Neurosci. *23*, 376–388. 10.1038/s41583-022-00582-9.
71. Cootes, T.F., Edwards, G.J., and Taylor, C.J. (1998). Active appearance models. In Computer Vision — ECCV'98, H. Burkhardt and B. Neumann, eds. (Springer Berlin Heidelberg), pp. 484–498.
72. Chang, L., Egger, B., Vetter, T., and Tsao, D.Y. (2021). Explaining face representation in the primate brain using different computational models. Curr. Biol. *31*, 2785-2795.e4. 10.1016/j.cub.2021.04.014.
73. Kietzmann, T.C., Swisher,J.D., König, P., and Tong, F. (2012). Prevalence of Selectivity for Mirror-Symmetric Views of Faces in the Ventral and Dorsal Visual Pathways. J. Neurosci. *32*, 11763. 10.1523/JNEUROSCI.0126-12.2012.
74. Tai Sing Lee (1996). Image representation using 2D Gabor wavelets. IEEE Trans. Pattern Anal. Mach. Intell. *18*, 959–971. 10.1109/34.541406.
75. Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355. 10.1038/nature06713.
76. Raghunathan, T.E., Rosenthal, R., and Rubin, D.B. (1996). Comparing correlated but nonoverlapping correlations. Psychol. Methods *1*, 178. 10.1037/1082-989X.1.2.178
77. Fisher, R.A. (1970). Statistical methods for research workers. In Breakthroughs in statistics: Methodology and distribution (Springer), pp. 66–70. 10.1007/978-1-4612-4380-9_6
78. Henin, S., Turk-Browne, N.B., Friedman, D., Liu, A., Dugan, P., Flinker, A., Doyle, W., Devinsky, O., and Melloni, L. Learning hierarchical sequence representations across human cortex and hippocampus. Sci. Adv. *7*, eabc4530. 10.1126/sciadv.abc4530.
79. Sherman, B.E., Graves, K.N., Huberdeau, D.M., Quraishi, I.H., Damisah, E.C., and Turk-Browne, N.B. (2022). Temporal Dynamics of Competition between Statistical Learning and Episodic Memory in Intracranial Recordings of Human Visual Cortex. J. Neurosci. *42*, 9053. 10.1523/JNEUROSCI.0708-22.2022.
80. Tallon-Baudry, C., Bertrand, O., Delpuech, C., and Pernier, J. (1996). Stimulus Specificity of Phase-Locked and Non-Phase-Locked 40 Hz Visual Responses in Human. J. Neurosci. *16*, 4240. 10.1523/JNEUROSCI.16-13-04240.1996.
81. Widmann, A., Schröger, E., and Maess, B. (2015). Digital filter design for electrophysiological data – a practical approach. Cut.-Edge EEG Methods *250*, 34–46. 10.1016/j.jneumeth.2014.08.002.