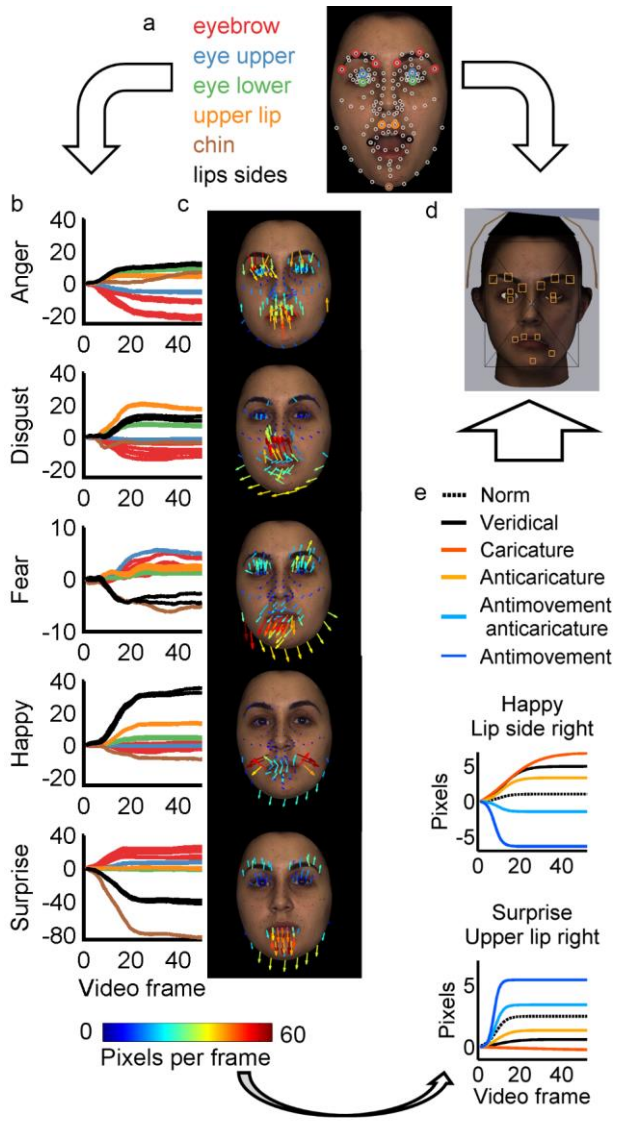

Figure 1. (a) We tracked 141 landmarks across frames, including 15 key landmarks (colored). (b) Plots of the vertical components of 15 key landmarks, averaged for each expression, indicate sigmoidal shapes. We derived veridical parameters by fitting logistic functions to normalized vertical displacements of key landmarks. (c) Quiver plots show landmark velocity. (d) We registered facial texture maps to a head model and animated key landmarks (yellow squares). (e) Movements are logistic functions reconstructed from logistic parameters "max" (asymptote), "slope" (bias) and "mid" (inflection point). We averaged veridical parameters over all movements and computed caricature levels by modifying distances of parameters from this "norm movement".

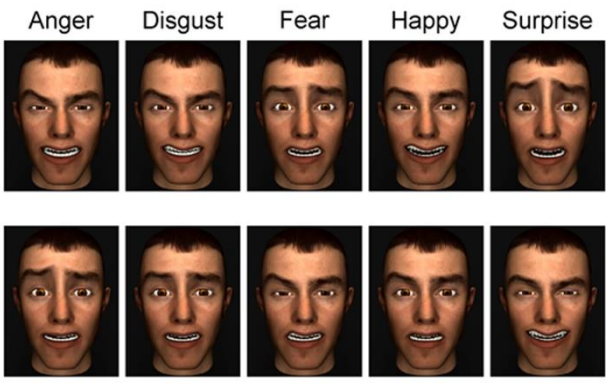

Figure 2. Caricatured expressions (final video frames) for one identity (top row) and those of the corresponding antimovements (bottom row).

Experimental procedures

For each of the 30 identity × expression combinations, we derived a new 4-label Type I index I sequence[8] to counterbalance the four caricature levels (caricature, anticaricature, antimovement anticaricature and antimovement), null events (blank screen) and target events (small white plus sign). In response to target events, participants pressed a response pad key with their right index fingers. All events appeared against a black screen for 200 ms followed by a blank screen for 500 msec. These sequences were concatenated in a pseudorandom order and the total sequence was divided into five scanning runs of 216 trials each. A structural scan was taken after the third session and

localizer scans were run after the second and fourth sessions. We designed localizer runs according to previously-published methods[13] we identified regions of interest (ROIs) that were face-selective (bilateral OFA and FFA), motion-sensitive (bilateral V5), and the face motion sensitive (right STS). Localizer runs presented 11 s blocks, each comprising one of four stimulus types: dynamic face or object videos or static face or object images (the final frames taken from the videos). Face videos[14] included eight identities (half female), presented in greyscale and transitioning from a neutral expression to either disgust, fearful, happy or sad. Each block presented all eight identities, with each expression presented twice. Dynamic objects[15] included spinning globes, ceiling fans, machinery, a candle and plants moving in wind. Participants identified via button press when a fixation dot located in the center of each movie turned from white to red on one-third of trials.

fMRI data acquisition, preprocessing and analysis

Data were collected on the 3T Allegra scanner (Siemens, Munich, Germany) at Royal Holloway, University of London. Volumetric data included a 1 mm^3 spatial resolution MPRAGE for each participant. Echo-planar volumes (56 slices, 2 mm^3 voxels, TR = 1.2 s, TE = 36.8 ms, FA = 30°) were collected using a high-resolution multi-band sequence. Scans were slice-time corrected, realigned, coregistered to anatomic scans, spatially-normalized to the Montreal Neurological Institute (MNI) standardized space and smoothed to 5 mm^3 full-width half maximum using conventional procedures in SPM12 (Wellcome Trust Centre for Neuroimaging, UCL, UK). Each first-level, individual-participant, general linear model (GLM) described below used AR(1) correction and a 128 ms high-pass filter. We computed contrasts in first-level GLMs then tested them for significance at the group level using mass-univariate second-level one-sample *t*-tests. Clusters were identified at $P < 0.001$ uncorrected and reported when also significant at a $P < 0.05$ cluster-level family-wise error rate.

RESULTS

We detected a cluster of direct effects (Fig. 3a and 3d) in early visual cortex (EVC), near lingual gyrus (Brodmann areas 18 and 19) and posterior cingulate (Brodmann area 29) with peak 12 -46 6 *MNI*. Another cluster appeared in anterior and orbitofrontal cortex (OFC) (Brodmann areas 10 and 11) and anterior cingulate (Brodmann area 32) with peak -14 42 -10 *MNI*. We aimed to measure variation in deviations from the norm so we statistically controlled for nuisance variability in the parameters' absolute magnitudes. We also examined functionally-defined regions of interest (ROIs), defined in individual participants using our localizer scans. No significant direct effects were observed in any localizer ROI, even for one-tailed *t*-tests at $P < 0.05$ uncorrected. Exploratory

testing also revealed widespread "familiarity" effects, in which fMRI responses were greater for familiar (caricatures, anticaricatures) than unfamiliar movements (antimovements, antimovement anticaricatures). These were observed in bilateral lateral and ventral occipito-temporal cortices with peaks 48 -62 2 and -44 -70 -2 *MNI* and right inferior frontal gyrus (Brodmann area 6) with peak 46 2 36 *MNI* (Fig. 3b) and in all localizer ROIs (Fig. 3c).

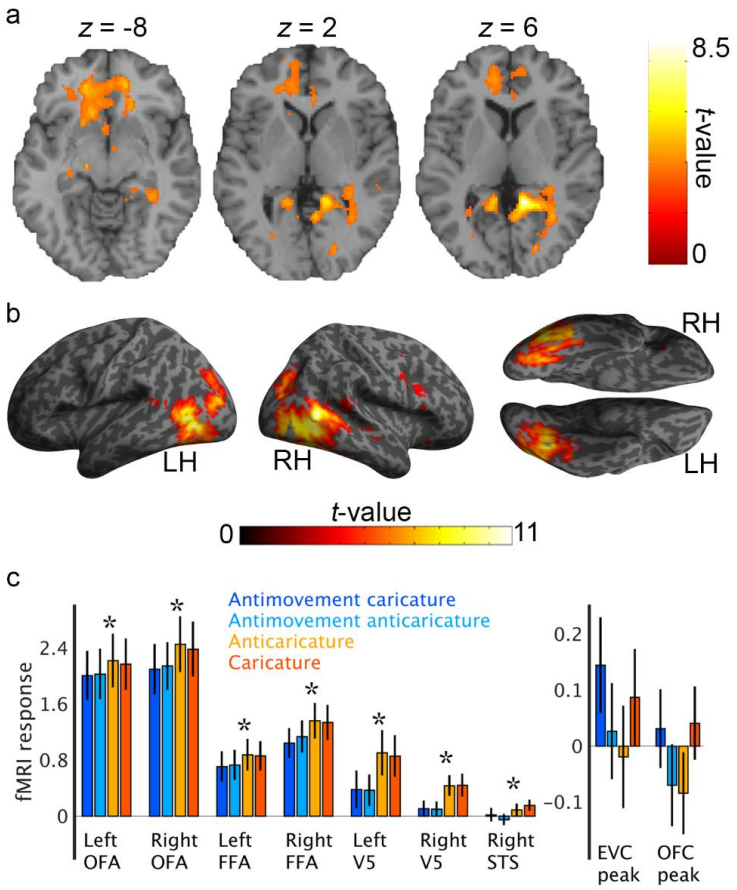

Figure 3. fMRI response amplitude effects of caricature level and familiarity. (a) Heightened responses to faces with movement parameters more distant from the norm movement (i.e., direct effects) exhibited peak effects in orbitofrontal (OFC) and early visual (EVC) cortices. (b) Statistical parametric maps ($P < 0.001$ uncorrected) showing heightened occipitotemporal responses; LH = Left hemisphere; RH = right hemisphere. (c) The left graph shows heightened responses to familiar expressions in regions of interest (ROIs) from the localizer scan (left panel). Stars represent significant two-tailed *t*-tests comparing mean responses to familiar and unfamiliar expressions at $P < 0.05$, Bonferroni-corrected for seven ROIs. The right panel shows caricature level means for the OFC and EVC peaks from (a), plotted for illustrative purposes. Error bars represent 95% confidence intervals.

The dissimilarity organization of these caricatures can also be used to identify "repetition suppression", a type of carryover effect[7]. Neurons selectively tuned to a stimulus dimension produce suppressed fMRI signals when repeated stimuli are similar along the tuned dimension[8]. This way, we identified regions that responded to dissimilarity between caricature levels of faces and their predecessors. The pattern of mass-univariate effects (Fig. 4a) appeared similar to that observed for familiarity (Fig. 3b), including bilateral lateral and ventral occipitotemporal areas with peaks at 54 -60 -4 and -30 -62 -14 *MNI* and right inferior frontal gyrus (Brodmann area 9) with peak 46 6 28 *MNI*. Because the largest carryover dissimilarity always entailed an expression change (caricatures versus anti-movements) and the smallest carryover dissimilarity always repeated an expression, we divided trials into within-expression and between-expression repetitions and verified that repetition suppression could survive for caricature

level dissimilarity in both cases. This was the case for ventral aspects of bilateral occipitotemporal cortex (Fig. 4b and 4c) and the localizer face-selective ROIs right occipital face (OFA) and fusiform face (FFA) areas (Fig. 4d). This suppression pattern appears to match that of the similarity structure we built into our caricatures (Fig. 4e), suggesting neurons tuned to approximately the same movement dimensions.

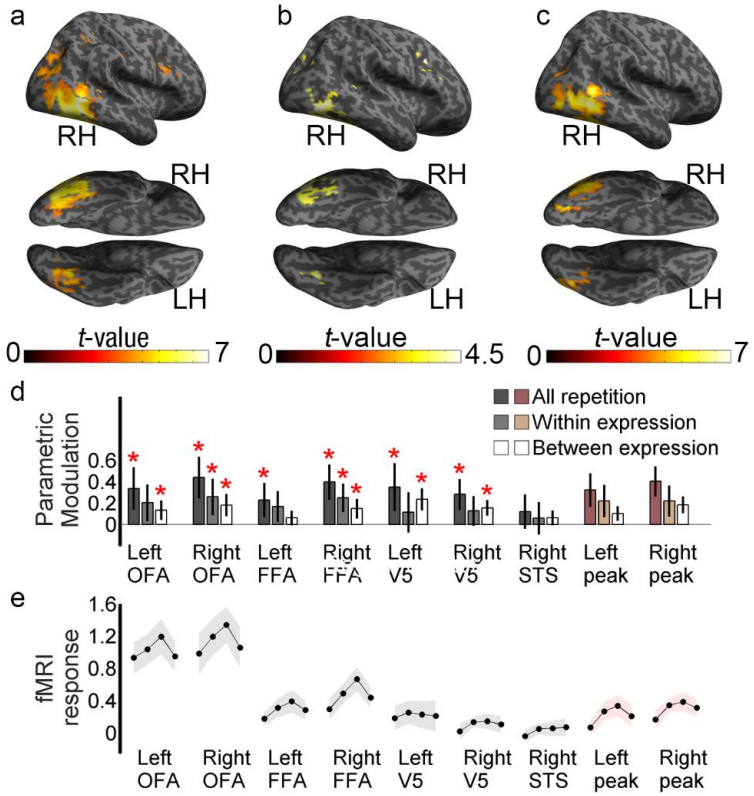

Figure 4. (a) Statistical parametric maps (P < 0.001 uncorrected) show parametric modulation by dissimilarity between caricature levels of faces and their predecessors for both (a) all repeated faces (b) within-expression repetitions (c) between-expression repetitions; LH = Left hemisphere; RH = right hemisphere. (d) Mean parametric modulation. For localizer ROIs (grey bars), stars represent one-tailed *t*-tests where parametric modulation was greater than zero, Bonferroni corrected. Pink bars represent mean values from 5 mm radius spheres around occipitotemporal clusters peak from (b), plotted for illustrative purposes. Error bars represent 95% confidence intervals over participants. (e) Mean responses to the five increasing dissimilarity values are plotted from left to right for each ROI, exhibiting increased fMRI responses with dissimilarity from preceding face. Shaded areas represent 95% confidence intervals.

Our movement-parameterized stimuli provided quantitative estimates of different motion dimensions (max, slope and mod). Over faces, each of these movement dimensions varies in its differences from its mean "norm" parameter (max norm, slope norm, mid norm). To identify which of these dimensions might be implemented in a norm-based coding scheme, we tested whether these averaged norm-based differences covaried with fMRI responses, with each predictor variable statistically controlling for both norm-based variation in other parameters and for variability in absolute magnitude (max mag, slope mag, mid mag). fMRI responses in primarily right occipitotemporal cortex were heightened when average landmark transitions were slower (slope norm) or earlier (slope mid) than average. There was a negative relationship between fMRI responses and slope norm (Fig. 5a) in right occipitotemporal cortex, peaking in lingual gyrus (Brodmann area 18) 26 -76 -6 *MNI* and extending laterally into inferior temporal gyrus. We found another negative relationship with fMRI responses, but for mid norm (Fig. 5b), in the right hemisphere peaking in

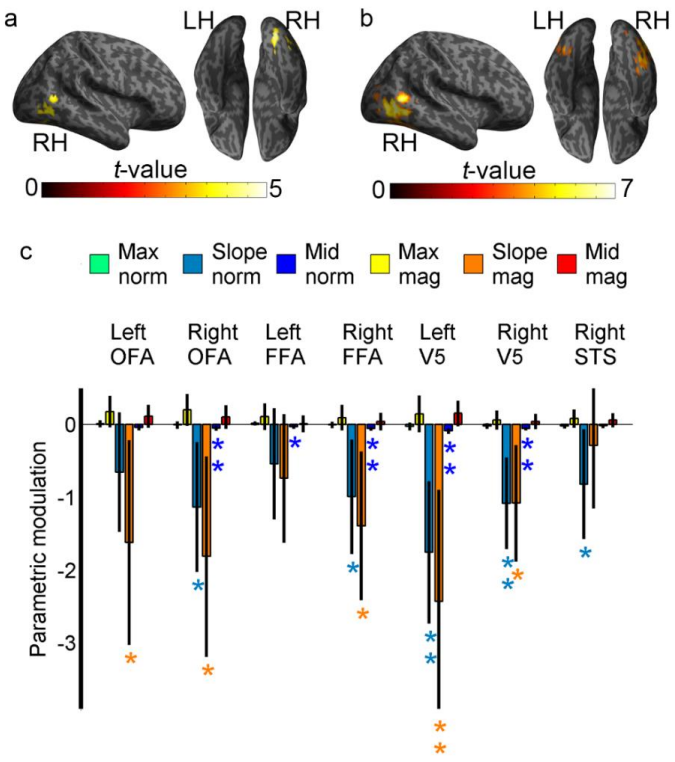

Figure 5. In the upper panels (a and b), statistical parametric maps ($P < 0.001$ uncorrected) show negative occipitotemporal parametric modulation by (covariation with): (a) differences in the slope parameter from the mean slope value (slope norm) and (b) differences in the mid parameter from the mean mid value (mid norm); LH = Left hemisphere; RH = right hemisphere. (c) Mean parametric modulation by deviations from the average parameter (norm) or absolute magnitude (mag) for the three logistic parameters (max, slope and mid) in localizer ROIs. Stars indicate parametric modulations significantly different from zero (two-tailed *t*-tests) at $P < 0.05$, Bonferroni-corrected for 11 ROIs. Error bars represent 95% confidence intervals

inferior temporal gyrus 46 -68 -2 *MNI* and extending ventrally into fusiform gyrus, and in the left hemisphere peaking in inferior occipital cortex (Brodmann area 19) at 46 -70 4 *MNI* and extending into middle temporal gyrus. We did not find significant mass-univariate effects for any of the other motion dimensions (norm max, max mag, slope mag, mid mag). Significant mass-univariate effects appeared nearby the V5 location, where motion-sensitivity (dynamic versus static stimuli) peaked at the group level at 46 -60 4 *MNI* in the right hemisphere. This result is consistent with the significant slope norm and mid norm results we found in our individual-participant localizer V5 ROIs (Fig. 5c). Other ROIs also showed small, but reliable, modulations of mid norm in right OFA and FFA (Fig. 5c).

RSA provided further evidence for regions implementing a similarity space isomorphic with our stimulus set. We implemented RSA using CoSMoMVPA[16] and searchlight mapping[17] using 8 mm radius spheres centered on every voxel. The matrix of caricature dissimilarity showed strong correlations with response pattern similarity throughout widespread regions of occipitotemporal cortex (Fig. 6a), and in every localizer ROI (Fig. 6b), suggesting that their visual codes match features of the similarity structure built into the caricatures.

We implemented dynamic causal modeling[18] to model how interactions among hierarchical levels (early visual, occipitotemporal and frontal) give rise to direct (Fig. 3) and carryover (Fig. 4) effects. Following the principle that modulation of a connection by an experimental condition alters how its downstream regions respond in that condition, we built models (Fig. 7) capable of explaining direct effects in EVC and OFC and carryover effects in bilateral occipitotemporal cortex (OCT). A comparison of their model evidences using Bayesian model comparison[19] showed a

posterior probability of nearly 1 favoring a model where direct effects depend on modulation in reciprocal connections between EVC and OFC, while carryover effects depend on bottom-up modulation from EVC.

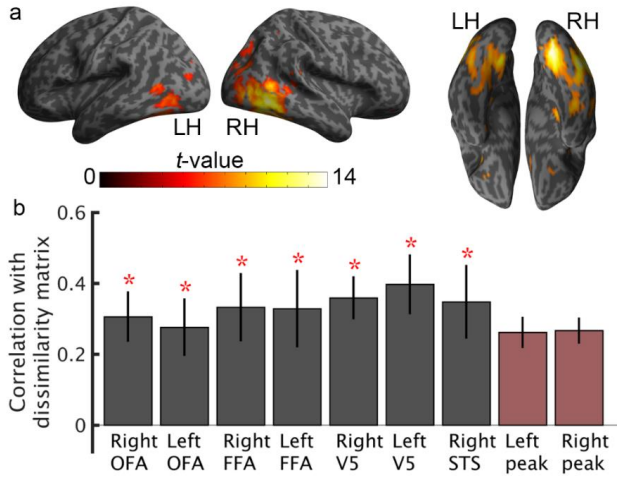

Figure 6. (a) Statistical parametric maps ($P < 0.001$ family-wise error corrected) showing lateral and ventral occipitotemporal voxels where local multivoxel response patterns correlated with similarity between caricature levels; LH = Left hemisphere; RH = right hemisphere. (b) Mean correlations between facial movement similarity and response pattern dissimilarity in localizer ROIs (gray) and 5 mm radius spheres surrounding peak voxels from (a) (pink). Error bars represent 95% confidence intervals.

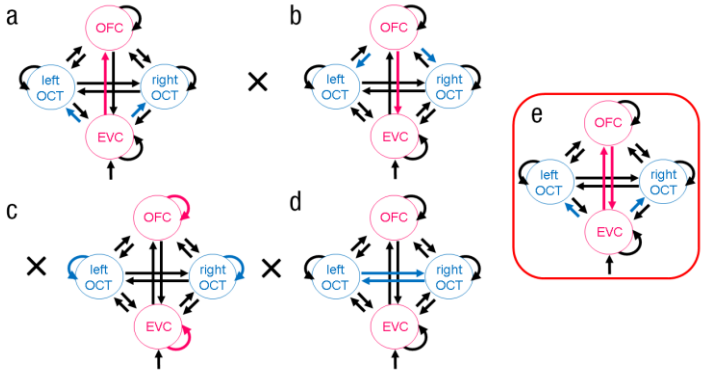

Figure 7. We tested dynamic causal models (DCMs) that included regions showing direct effects (Fig. 1a) in EVC and OFC (magenta circles) and carryover effects (Fig. 2b) in bilateral occipitotemporal cortex (OCT) (cyan circles). All models were fully intrinsically connected (black arrows between regions) with all face stimuli as input in EVC. Our model space attempted to explain direct and carryover effects in these regions by varying which connections were modulated by caricature level (connections shown in cyan) or carryover effects (connections shown in magenta). The DMS could modulate: (a) forward connections, (b) backward connections, (c) self connections, (d) lateral connections (carryover modulation only). (e) The highest-evidence model (posterior probability near 1.0) exhibited both forward and backward modulated connections.

DISCUSSION

When experiencing caricatured facial dynamics, the brain instantiates similarity space representations, for which differences from a norm might play a role. These representations manifest in different ways at different levels of an apparent hierarchy. Early visual cortex showed direct effects of caricature (similarity to norm). Ostensibly higher levels of the visual hierarchy, in ventral and lateral regions of occipitotemporal cortex did not show direct effects, but instead manifested sensitivity to the similarity relationships among our caricatured stimuli in two ways. First, they exhibited repetition suppression by caricature level similarity. Second, their multivariate response patterns were inter-related in a way that matched the caricature similarity of our stimulus set. This finding accords with other studies of high-level vision, finding various visual similarity spaces at approximately this level of the visual system[20]. The similarity spaces we detected were found across specialized functional areas including face-selective areas OFA and FFA and the motion-sensitive area V5. Bilateral V5 may participate in norm-based coding of the speed (slope) and

time (mid-point) of movements. At an even higher level of this putative hierarchy, in OFC and rostral anterior cingulate, direct effects re-emerged. Having established these three hierarchical levels, we tested for interactions between them and found that reciprocal OFC-EVC connectivity gives rise to direct effects, and feedforward connectivity from EVC drives repetition suppression in downstream occipitotemporal areas.

To date, no studies have tested if norm-based coding supports the challenging computational problem of perceiving complex, multidimensional and dynamic stimuli. Theories of social and neural processing should not continue to be based only on studies with static images, which do not challenge visual cortex, as stimuli "in the wild" would. Our findings shift the emphasis away from simply localizing face-selective responses and toward understanding the nature of representations and the hierarchical mechanisms that give rise to them.

REFERENCES

1. Valentine T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q J Exp Psychol A* 43:161-204.

2. Light LL, Kayra-Stuart F, Hollander S. (1979). Recognition memory for typical and unusual faces. *J Exp Pschol Hum Learn* 5:212-28.

3. Blanz V, O'Toole AJ, Vetter T & Wild H. (2000). On the other side of the mean: the perception of dissimilarity in human faces. *Perception* 29:885-891.

4. Freiwald WA, Tsao DY, Livingstone MS. (2009). A face feature space in the macaque temporal lobe. *Nat Neurosci* 12:1187-96.

5. Leopold DA, Bondar IV, Giese MA. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* 442:572-5.

6. Loffler G, Yourganov G, Wilkinson F, Wilson HR. 2005. fMRI evidence for the neural representation of faces. *Nat Neurosci* 8:1386-90.

7. Aguirre GK. (2007). Continuous carry-over designs for fMRI. *Neuroimage* 35:1480-94.

8. Grill-Spector K, Henson R, Martin A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn Sci* 10:14-23.

9. Kriegeskorte N, Kievit RA. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci* 17:401-12.

10. Furl, N., Hadj-Bouziane, F., Liu, N., Averbeck, B.B., Ungerleider, L.G. Dynamic and static facial expressions decoded from motion-sensitive areas in the macaque monkey. J. Neurosci. **32**, 15952-15962 (2012).

11. Yin L, Chen X, Sun Y, Worm T, Reale M, 2008. A high-resolution 3D dynamic facial expression database. In: *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*.

12. Chen, J., Tiddeman, B., 2010. Multi-cue facial feature detection and tracking under various illuminations. *Int J Robot Autom* 25, 162–171.

13. Furl N, Henson RN, Friston KJ, Calder AJ. (2013). Top-down control of visual responses to fear by the amygdala. *J Neurosci* 33:17435-43.

14. Van der Schalk, J., Hawk, S.T., Fischer, A.H. & Doosje, B.J. Moving faces, looking places: The Amsterdam Dynamic Facial Expressions Set (ADFES). *Emotion* 11, 907–920 (2011).

15. Fox, C.J., Iaria, G. & Barton, J.J. Defining the face processing network: optimization of the functional localizer in fMRI. *Hum Brain Mapp* 30, 1637–1651 (2009).

16. Oosterhof, N.N., Connolly, A.C. & Haxby J.V. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab / GNU Octave. Front. *Neuroinform* 10, 27 (2016).

17. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103, 3863-8 (2006).

18. Friston, K.J., Harrison, L. & Penny, W. (2003). Dynamic causal modeling. *Neuroimage* 19:1273–1302.

19. Penny WD, Stephan KE, Mechelli A, Friston KJ. (2004). Comparing dynamic causal models. *Neuroimage* **22**, 1157–1172.

20. Carlin JD, Kriegeskorte N. (2017). Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Comput Biol* 13:e1005604.