nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-019-0625-3

Modelling face memory reveals task-generalizable representations

Jiayu Zhan[1], Oliver G. B. Garrod[1], Nicola van Rijsbergen[1] and Philippe G. Schyns[1,2*]

Current cognitive theories are cast in terms of information-processing mechanisms that use mental representations[1–4]. For example, people use their mental representations to identify familiar faces under various conditions of pose, illumination and ageing, or to draw resemblance between family members. Yet, the actual information contents of these representations are rarely characterized, which hinders knowledge of the mechanisms that use them. Here, we modelled the three-dimensional representational contents of 4 faces that were familiar to 14 participants as work colleagues. The representational contents were created by reverse-correlating identity information generated on each trial with judgements of the face's similarity to the individual participant's memory of this face. In a second study, testing new participants, we demonstrated the validity of the modelled contents using everyday face tasks that generalize identity judgements to new viewpoints, age and sex. Our work highlights that such models of mental representations are critical to understanding generalization behaviour and its underlying information-processing mechanisms.

The cognitive mechanism of recognition is guided by mental representations that are stored in memory[1–4]. Personal familiarity with faces (for example, as family members, friends or work colleagues) provides a compelling everyday illustration because the information contents representing familiar faces in memory must be sufficiently detailed to enable accurate recognition (that is, identifying 'Mary' among other people) and sufficiently versatile to enable recognition across diverse common tasks—for example, identifying Mary in different poses, at different ages or identifying her brother based on family resemblance[5–7]. And yet, it remains a fundamental challenge to reverse engineer the participants' memory to model and thereby understand the detailed contents of their representations of familiar faces. This challenge is a cornerstone of understanding the brain mechanisms of face identification, because they process the contents to predict the appearance of the familiar face of 'Mary' in the visual array and to selectively extract its identity information to generalize behaviour across common tasks.

We studied how our own work colleagues recognize the faces of other colleagues from memory. The work environment provides a naturally occurring and common medium of social interactions for all participants, who had a minimum of six months exposure to the people whose faces the study tested. To model the three-dimensional (3D) face-identity information stored in their memory, we developed a methodology based on reverse correlation (see Fig. 1a and 'Reverse-correlation experiment' in Methods) and a generative model of 3D face identity (GMF, see Fig. 1b and 'Generative model of face identity' in Methods), separately for 3D shape and 2D texture information (see Supplementary Fig. 1a for 3D face parameters).

On each experimental trial, our GMF synthesized a set of six new 3D faces (see 'random faces' in Fig. 1a), each with a unique and randomly generated identity. Critically, each face shared other categorical face information (that is, sex, age and ethnicity) with one of the 4 faces that were personally familiar to each one of our 14 participants as work colleagues—for example, the familiar target face of 'Mary'. To achieve this, we used a general linear model (GLM) to decompose the familiar target face into a categorical component (for example, for 'Mary', the average of all white female faces of 30 years of age) plus a residual component that defines the specific identity of the familiar face (see 'identity modelling' in Fig. 1b). We then generated new random identities by keeping the categorical component of the target constant (for example, white female, 30 years of age) and adding a random component of identity (see 'identity generation' in Fig. 1b and 'Reverse-correlation experiment' and 'Random face identities' in Methods for details). Participants saw these randomly generated faces in full frontal view and selected the one that most resembled the familiar target (for example, 'Mary') and rated its similarity to the target on a six-point Likert scale, ranging from not at all (1) to highly similar (6). To resolve the task, participants must compare the randomly generated faces presented on each trial with their mental representation of the familiar target in full frontal view. Therefore, each face selected comprises a match to the participant's mental representation of the target, which is estimated by the similarity rating of that face.

After many such trials, we used reverse correlation[8] to estimate the information content of the mental representation of each target familiar face ($n = 4$, see Supplementary Fig. 1b) in each participant ($n = 14$, see 'Reverse-correlation experiment' in Methods). Specifically, we built a statistical relationship between the information content of the faces that the participant selected on each trial with their corresponding similarity ratings. In a second stage, we tested with a new group of participants ($n = 12$; the validators; see 'Generalization experiments' in Methods for details) whether these modelled mental representations were sufficiently detailed to enable identification of each target familiar face and sufficiently versatile to enable resemblance judgements across diverse everyday tasks—that is, generalization across new viewpoints, age and siblings.

To reconstruct the information contents of mental representations, we used linear regression to compute the single-trial relationship between <similarity ratings, random face-identity components> for each target familiar face and participant. Specifically, we computed separate regressions between the similarity ratings and each 3D shape vertex and each RGB texture pixel that comprised the face-identity components. We then used the resulting Beta coefficients to model the 3D shape and texture identity components that characterize the participant's mental representation of each familiar face in the GMF (see Supplementary Fig. 2 and 'Analyses', 'Linear

[1]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK. [2]School of Psychology, University of Glasgow, Glasgow, UK.
*e-mail: philippe.schyns@glasgow.ac.uk

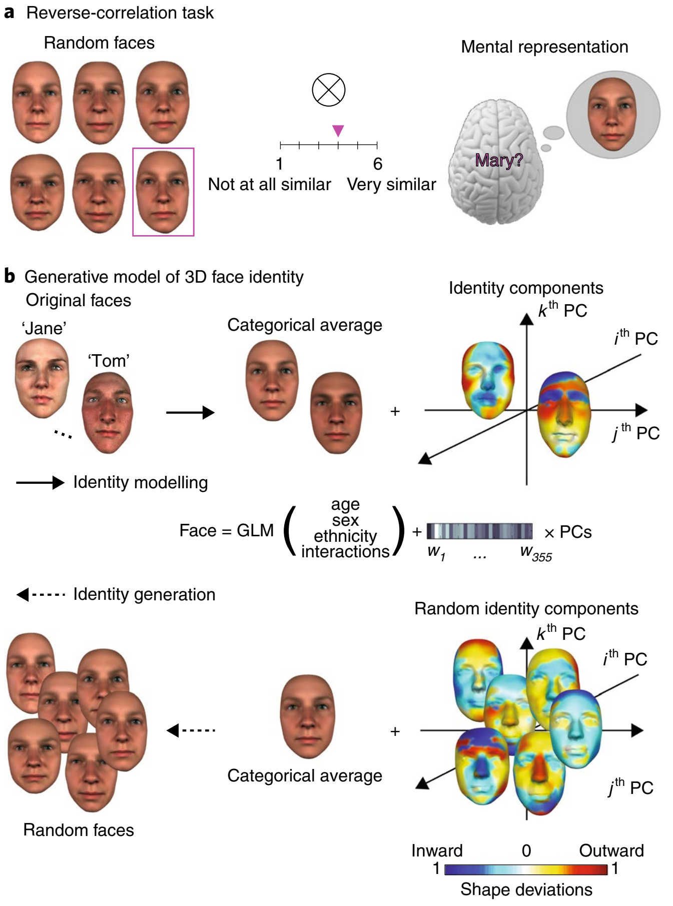

Fig. 1 | Reverse-correlating mental representations of familiar faces. **a**, Task. Illustrative experimental trial with six randomly generated face identities. We instructed participants to use their memory to select the face most similar to a familiar identity (here, 'Mary') and then to rate the similarity of the selected face (purple frame) to their memory of 'Mary' (purple pointer). **b**, The GMF. In its forward computation flow (see identity modelling, solid arrow), the GLM decomposes a 3D, textured face (for example, 'Jane' or 'Tom') into a non-identity face-shape average capturing the categorical factors of face sex, ethnicity, age and their interactions plus a separate component that defines the identity of the face (illustrated by the 3D shape decomposition; 2D texture, not illustrated, is independently and similarly decomposed). Heat maps indicate the 3D shape deviations that define 'Jane' and 'Tom' in the GMF (as w_i weights that multiply the principal components that control the residual identity information) in relation to their categorical averages. In the reverse flow (see dashed arrow of identity generation), we can randomize the 3D shape identity component (and 2D texture component, not illustrated here), add the categorical average of 'Jane' (or 'Tom') and generate random faces, each with a unique identity that share all other categorical face information with 'Jane' and 'Tom'. PC, principal component.

regression model' and 'Reconstructing mental representations' in Methods for details).

With this approach, we can formally characterize and then compare the participant's mental representation of a familiar face with the ground truth face—that is, the objective identity component of the scanned familiar face (see Supplementary Fig. 1b). We focus only on 3D shape because there were very few and only non-systematic relationships for texture (see Supplementary Fig. 3). To illustrate the approach, grey faces on the *x* axis of Fig. 2a show the ground truth identity component of 'Mary' in the GMF for inward and outward 3D shape deviations in relation to the categorical average (that is, of all white females of 30 years of age, like 'Mary'). For example, Mary's nose is objectively thinner than the average of white females of her age, and so these vertices deviate inward (darker grey tones indicate increasing deviations). Likewise, her more pouty mouth is shown as an outward 3D shape deviation. The *y* axis of Fig. 2a uses the same format to show the mental representation of Mary in one typical participant, where colours indicate increasing deviations. These contents reveal faithful representations of, for example, a thinner nose and a pouty mouth (see 'Analyses' and 'Vertex contribution to mental representations' in Methods). A scatter plot visualizes the vertex-by-vertex fit between the mental representation (*y* axis) and the ground truth 3D face (*x* axis). The white diagonal line provides a veridical reference, where the identity component in the mental representation is identical to the ground truth face, for every single 3D vertex. This is because the mental representation and ground truth faces are both registered in the same space of 3D vertices[9].

Our analyses reveal the specific vertices near the veridical line that faithfully represent 'Mary' in the mind of this participant as coloured dots reported on the scatter and located on the *y*-axis faces in Fig. 2a. These vertices indicate faithful representations because they are significantly closer to the ground truth faces than a null distribution of representations arising from chance ($P < 0.05$, two-sided *t*-test, with a null distribution that iterated 1,000 times the analyses using a random permutation of the participant's choice responses on each iteration, see 'Analyses' and 'Vertex contribution to mental representation' in Methods for details). By contrast, white vertices away from the veridical line did not faithfully represent the identity. We repeated the analysis of represented contents for each participant ($n = 14$) and familiar face ($n = 4$). Figure 2b reports the collated group results, using the format of Fig. 2a, where colours now indicate n, that is, the number of participants who faithfully represented that identity in their mind with this particular 3D shape vertex. Figure 2b demonstrates that mental representations comprised similar information contents across the 14 individual participants. Most (10 of 14) faithfully represented 'Mary's' thin nose, 'John's' receding eyes and wider upper face (13 of 14), 'Peter's' prominent eyebrow and jawline (13 of 14), 'Stephany's' protruding mouth (13 of 14).

Such convergence of represented contents across participants suggests that the face representations could be multivariate (that is, comprising contiguous surface patches rather than isolated vertices). As a final step, we extracted the main multivariate components of represented surface patches. To this end, we applied across observers ($n = 14$) and familiar faces ($n = 4$) the non-negative matrix factorization (NNMF)[10] to the faithfully represented 3D vertices (see 'Analyses' and 'Components of memory representation' in Methods). Figure 3a shows the multivariate components that faithfully represent four target identities and Fig. 3b shows their combinations for the diagnostic components of each target identity (for example, for 'Mary', the red background heat map; for 'Stephany', the green one). Importantly, these diagnostic components of familiar-face identity have complementary non-diagnostic components (that is, the grey background heat maps in Fig. 3b), which capture variable face surfaces that do not comprise the participants' mental representations.

Here, we develop the critical demonstration that the information contents of the mental representations we modelled are valid. That is, the contents enable accurate identification of each target face and they also enable resemble tasks that preserve their identity. We asked a new group of participants (called 'validators') to resolve a variety of resemblance tasks that are akin to everyday tasks of face recognition. Success in these tasks would demonstrate that the diagnostic components derived from the previous experiment comprise identity information that can be used in different generalization tasks. Therefore, although the components are extracted under one viewpoint (full face), one age (for each identity) and one sex

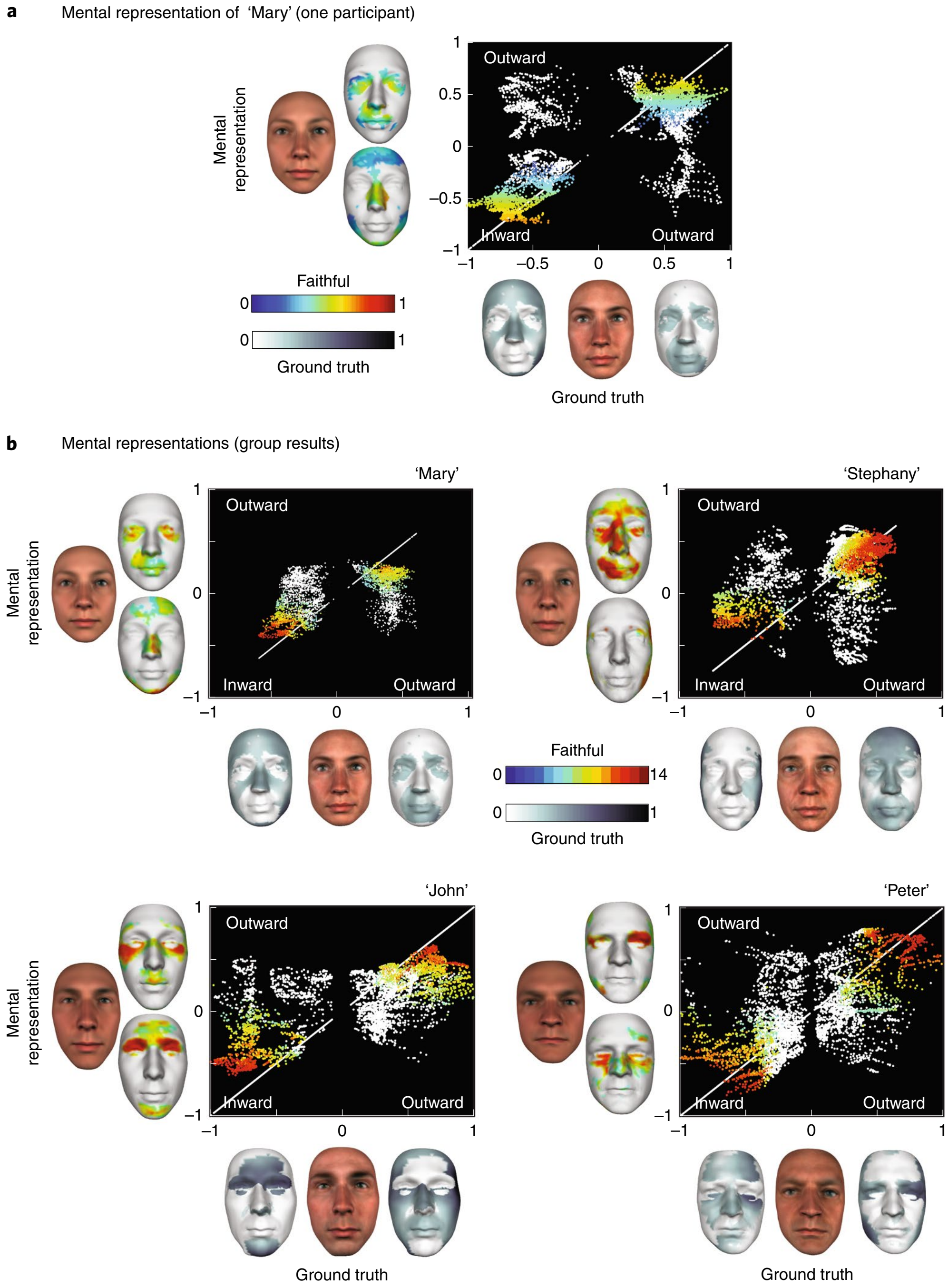

Fig. 2 | Contents of mental representations of familiar faces. **a**, Mental representation of 'Mary' (a typical participant). Ground truth: 3D vertex positions deviate both inward (–) and outward (+) from the categorical average to objectively define the shape of each familiar-face identity. Greyscale values reported on the flanking faces colour code the normalized magnitudes of inward and outward deviations from the categorical average. Mental representation: inward and outward coloured faces highlight the individual 3D vertices, the positions of which faithfully deviate from the categorical average in the GMF ($P < 0.05$, two-sided *t*-test). Blue to red colours represent the normalized magnitudes of their deviations. Two-dimensional scatter plots: scatter plots indicate the relationship between each vertex deviation in the ground truth (normalized scale on the *x* axis) and the corresponding vertex in the memory representation (normalized scale on the *y* axis). The white diagonal line provides the reference of veridical mental representation in the GMF—a hypothetical numerical correspondence between each shape vertex position in the ground truth face and in the mental representation of the same face. White dots indicate vertices that were not faithfully represented. **b**, Mental Representations (group results). As in Fig. 2a, except that the colour map now reflects the number of participants ($n = 14$) who faithfully represented this particular shape vertex.

(that of the identity), here we tested the generalization of identification performance to new viewpoints, ages and sex.

For this demonstration, we synthesized new diagnostic (as opposed to non-diagnostic) faces that were parametrically controlled for the relative strength of the diagnostic multivariate components of identity versus their non-diagnostic complement (see Fig. 4a and 'Generalization experiments' and 'Stimuli' in Methods). It is important to emphasize that both diagnostic and non-diagnostic faces are

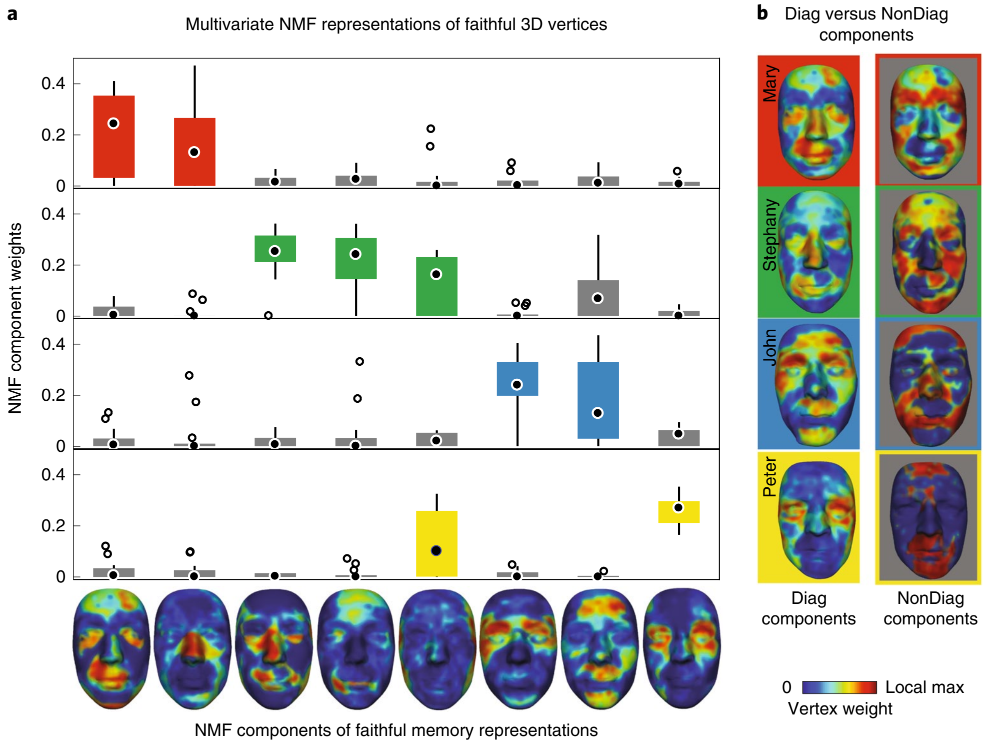

Fig. 3 | NNMF multivariate and compact representations. a, NNMF representations of faithful 3D vertices across the mental representations of participants. The *x*-axis heat map presents each NNMF component, where colours indicate the relative weight of each shape vertex in the component (normalized by maximum weight across components). Box plots on the *y* axis show the loading of each NNMF component on the faithful representations ($n=14$, one per participant) of each familiar identity ($n=4$ familiar identities), with coloured boxes indicating above-threshold (>0.1) loading for NNMF components. In box plots, the bottom (versus top) edges indicate the 25th (versus 75th) percentile of the distribution; the whiskers cover the +2.7 s.d.; the larger central circle indicates the median; the outliers are plotted in smaller circles outside the whiskers. **b**, Diagnostic (Diag) and non-diagnostic (NonDiag) components for each familiar identity. Heat maps in the left column show the diagnostic component for each familiar identity; heat maps in the right column show the complementary non-diagnostic components.

equally faithful representations of the original ground truth. That is, their shape features are equidistant from the shared categorical average. However, whereas the diagnostic components deviate from the average with multivariate information extracted from the participants' mental representations, the non-diagnostic components do not. We hypothesized that, though equidistant from the categorical average, only the diagnostic components will impact performance on the resemblance tasks. For all synthesized faces, we changed their viewpoint (rotation of −30°, 0° and +30° in depth), age (to 80 yr old) and sex (to opposite) using the generative model (see Supplementary Figs. 5–8 for each familiar target).

In three independent resemblance tasks—changes of viewpoint, age and sex—we tested the identification performance of 12 validators on the diagnostic and non-diagnostic faces using a five-alternative force choice task (that is, responding one of four familiar identities plus a 'don't know' response, see 'Generalization experiments' and 'Procedure' in Methods). In each task, for each identity we found a significantly higher identification performance for diagnostic faces (see Fig. 4b, red curves) than for non-diagnostic faces (black curves)—that is, a fixed effect of 'face type' in a mixed-effects linear model. For 'Mary', $F(1,12.76)=315.49$, $P<0.001$, estimated slope = 0.297, 95% confidence intervals = [0.264, 0.33]; for 'Stephany', $F(1,20.62)=25.068$, $P<0.001$, estimated slope = 0.058, 95% confidence intervals = [0.035, 0.081]; for 'John', $F(1,12)=21.369$, $P<0.001$, estimated slope = 0.143, 95% confidence intervals = [0.083, 0.204]; and for 'Peter', $F(1,12.01)=5.76$, $P=0.034$, estimated slope = 0.095, 95% confidence intervals = [0.017, 0.173] (see 'Generalization experiments' and 'Analyses' in Methods for the detailed specification and Supplementary Tables 3–6 for the full statistical analysis of the models). Thus, the diagnostic contents of the mental representations we modelled do indeed contain the information that can resolve identity and resemblance tasks.

Mental representations stored in memory are critical to guide the information-processing mechanisms of cognition. Here, with a methodology based on reverse correlation and a 3D face-information generator (that is, our 3D GMF), we modelled the information contents of mental representations of 4 familiar faces in 14 individual participants. We showed that the contents converged across participants on a set of multivariate features (that is, local and global surface patches) that faithfully represent 3D information that is objectively diagnostic of each familiar face. Critically, we showed that validators could identify new faces generated with these diagnostic representations across three resemblance tasks—that is, changes of pose, age and sex—but performed much worse with equally faithful, but non-diagnostic features. Together, our results demonstrate that the modelled representational contents were both sufficiently precise to enable face identification within task and versatile enough to generalize usage of the identity contents to other resemblance tasks.

At this stage, it worth stepping away from the results to emphasize that it is remarkable that the reverse-correlation methodology works at all, let alone produces robust generalization across resemblance tasks. In the experiment, we asked observers to rate the resemblance between a remembered familiar face, and randomly generated faces, that by construction are very unlike the target face (never identical, and almost never very similar). And yet, our results

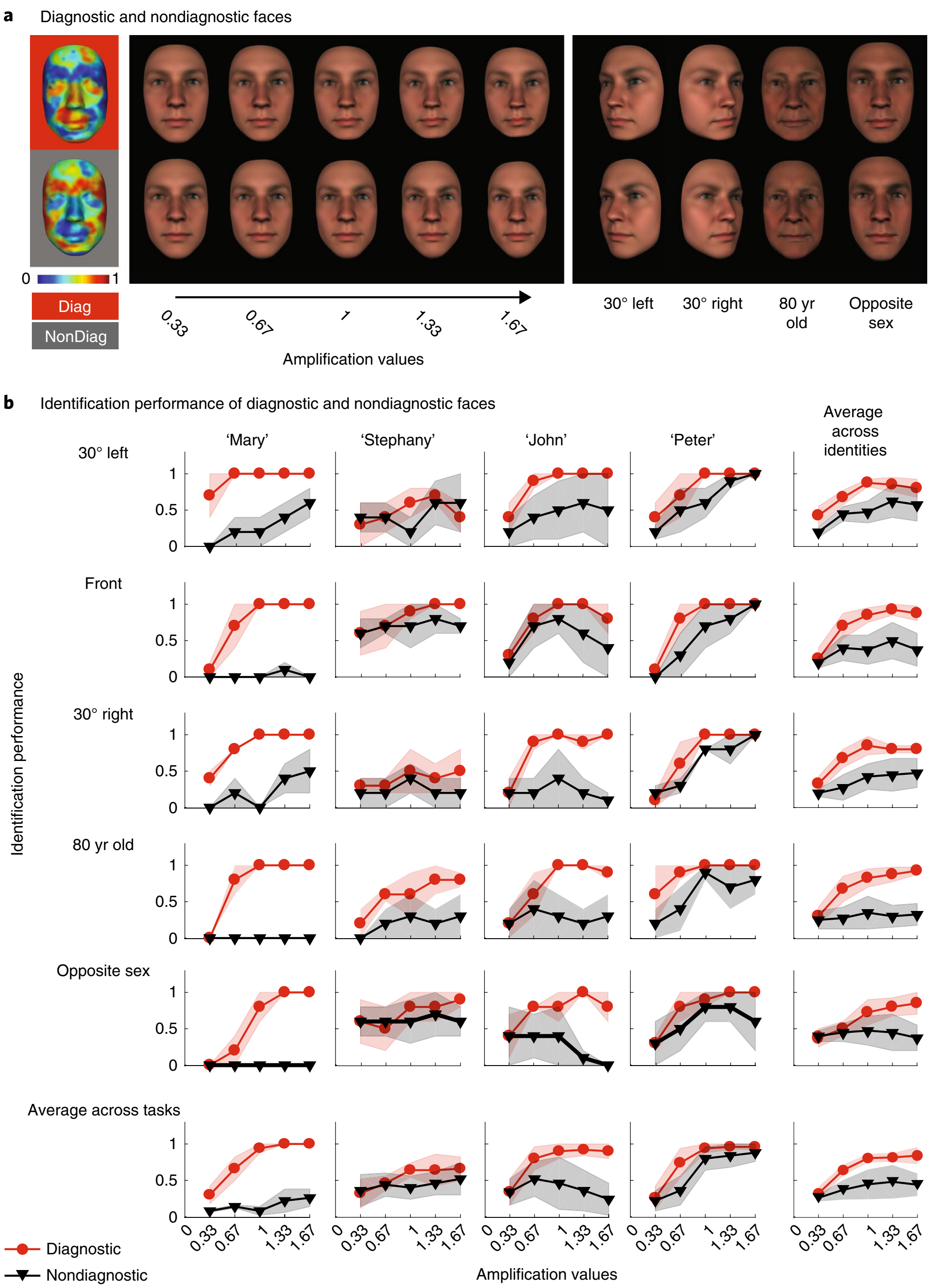

Fig. 4 | Generalization of performance across tasks. a, Diagnostic and non-diagnostic faces. Left: the red background map shows the multivariate diagnostic components of faithful 3D shape representation of 'Mary'; the grey background map shows the non-diagnostic complement (1 – diagnostic components). Middle: faces synthesized with increasing amplification (0.33 to 1.67) of the diagnostic (top) versus non-diagnostic (bottom) components. Right: for each synthesized face, we changed its viewpoint (30° left and 30° right), age (80 yr old) and sex, shown here for faces synthesized at amplification = 1. **b**, Task performance. For each condition of generalization (row) and familiar identity (column), 2D plots show the median identification performance computed across 12 validators (*y* axes) for faces synthesized with the diagnostic (red curves) and non-diagnostic (grey curves) faces, at different levels of amplification of the multivariate components (*x* axes). Shaded regions indicate median absolute deviations of identification performance.

show that the representational contents we modelled following such a task were in fact part of the contents that objectively (that is, faithfully) support identity recognition. This raises a number of important points that we now discuss.

There has been a recent surge of interest in modelling face representations from human memory[11–13]. These studies used 2D face images and applied dimensionality reduction (for example, principal component analysis (PCA)[14] and multidimensional scaling) to formalize an image-based face space, where each dimension is a 2D eigenface or classification image—that is, pixel-wise RGB (or L*a*b) colour-space values. To understand the contribution of each 2D face-space dimension to memory representations (including their neural coding), researchers modelled the relationship between projected weights of the original 2D face images on each dimension and participants' corresponding behavioural[13] (and brain[11,12]) responses.

These studies contributed important developments in face-identification research because they addressed the face-identity contents that the brain uses to guide face-identification mechanisms. Our aim was to model the face-identity contents in the generative 3D space of faces (not the 2D space of their image projections) and to use these models to generate identification information in resemblance tasks that test the generalizability of identity information. It is important to clarify that we modelled identity information in a face space that belongs to the broad class of 3D morphable, active appearance models (AAMs)[15,16] of facial synthesis . These models contain full 3D surface and 2D texture information about faces, and so with their better control superseded the former generation of 2D image-based face spaces[14,17,18]. To synthesize faces, we used our GMF to decompose each face identity as a linear combination of components of 3D shape and 2D texture added to a local average (that summarizes the categorical factor of age, gender, ethnicity and their interactions; compare with Fig. 1b). To model the mental representations of faces, we estimated the identity components of shape and texture from the memory of each observer. These components had generative capacity and we used them to precisely control the magnitude of identity information in new faces synthesized to demonstrate generalization across pose, age and sex. Thus, we used the same AAM framework for stimulus synthesis, mental representation estimation and generation of generalizable identities.

There is a well-known problem with using AAMs to model the psychology of face recognition. Perceptual expertise and familiarity are thought to involve representations of faces that enable the greater generalization performance that is widely reported[19–22]. However, AAMs typically adopt a brute-force approach to identity representation: a veridical (that is, totally faithful) deviation of each physical shape vertex and texture pixel from an average. Thus, as AAMs overfit identity information, they appear as a priori weak candidate models to represent perceptual expertise with faces[18]. Our approach of studying the contents of mental representations suggests a solution to this conundrum. We showed that each observer faithfully represented only a proportion of the objective identity information that defines a familiar-face identity. Our key theoretical contribution to face space is to formalize the subjective 3D diagnostic information as a reduced set of multivariate face features that can be construed as dimensions of the observer's face space. Observers develop these dimensions when they interact with the objective information that represents a new face identity in the real world. We modelled the objective information that is available to the observer for developing their face-space dimensions via learning as the veridical shape and texture information of the AAM[18,23,24]. Key to demonstrating the psychological relevance of our psychological 3D face-space dimensions is that they should comprise identity information that is sufficiently detailed to enable accurate face identification and sufficiently versatile to enable similarity judgements of identity in resemblance tasks. We demonstrated this potential when validators identified faces synthesized with the diagnostic dimensions in novel resemblance tasks. Thus, by introducing reduced faithful mental representations of identity information in the objective representations of AAMs we provide the means of modelling the subjective psychological dimensions of an individual's face space.

Our work could be extended to precisely track the development of the psychological dimensions of face space if we tasked observers with learning new identities (an everyday perceptual expertise task[18,25]). Our AAMs enable a tight control of objective face information at synthesis, such as ambient factors of illumination, pose and scale, but also categorical factors of gender, sex, age and ethnicity, and components of identity. Thus, we could tightly control the statistics of exposure to faces in individual observers (and even orthogonalize them across observers), and model and compare the diagnostic dimensions of the psychological face space that are learned, and finally test their efficacy as we did here. Then, when we understand how ambient and categorical factors influence performance as a function of differential perceptual learning, we can switch to understanding familiar-face identification in the wild, by progressively introducing simulations of ambient factors (for example, identifying the face of someone walking by a street lamp at night) and observe their specific effects on performance (for example, ambient changes in face size, shading and cast shadows). Otherwise, all ambient and categorical factors remain naturally mixed up, and the influence of each factor to identification performance becomes nearly impossible to disentangle, precluding a detailed information-processing understanding of face-identification mechanisms.

Our results suggest that human observers use face-shape information over texture to represent familiar identities. At this stage, it is important to clarify that shape and texture have different meanings in different literatures. For example, some authors in psychology discuss shape-free faces when referring to 2D images synthesized by warping an identity-specific texture to an identical 'face shape' (defined as a unique and standard set of 2D coordinates that locate a few face features[26]). However, it is important to emphasize that the warped textures are not free of 3D shape information (for example, that which can be extracted from shading[27]). In computer graphics, the generative model of a face comprises a 3D shape per identity (here, specified with 4,735 3D vertex coordinates), lighting sources (here, $n=4$) and a shading model (here, Phong shading[28]). The shading model interacts with shape and texture to render the 3D face as a 2D image. To illustrate the effects of this rendering, Supplementary Fig. 9 shows how applying the same 2D textures (rows) to different 3D face shapes (columns) generates 2D images with different identities. We used the better control afforded by computer graphics to generate our face images and found that shaded familiar-face shape was more prevalent in the face memory of individual participants than face texture.

A general question with reverse-correlation tasks is whether the resulting models represent a particular visual category (here, the visual identity of a face) or the task from which the model was reconstructed[24,29–31]. We contributed to this debate by showing that the identity information reconstructed in one task had efficacy in other tasks that involved identity. Importantly, the tasks were designed to test two classes of factors: ambient and categorical. For example, we showed that the identity component extracted in one ambient viewpoint (full face, 0°) could be used to generalize identification of the same face under two new ambient viewpoints (−30° and +30° of rotation in depth). We also showed that the identity component extracted for identities (all less than 40 yr of age) generalized to older age (80 yr). Furthermore, we also showed that although extracted from a given sex, the identity component would generalize to another sex, a kinship task. Hence, we found no marked differences due to the effect of task of extraction of the identity component. Rather, the extracted representational basis is useful for all tasks tested, whether using ambient or categorical factors of face variance. This therefore suggests that we have tapped into some essential information about familiar-face representation. However, we acknowledge that the generalizations we observe might still be a function of an interaction between the nature of memory and the similarity task from which we estimated the identity component. The component could have differed had the task been more visual than memory based (for example, identification of the same face under different orientations or a visual matching task) and we might not have derived an identity component that enabled such effective generalization. In any case, the memorized identity components that enable task generalization reflect an interaction between memory and the input information available to represent this identity[24,32]. Observers can compare this memory representation for that identity with a representation of the visual input for successful identification.

Our models of mental representation should be construed as the abstract information goals (that is, the contents) that the visual system predicts when identifying familiar faces. We call them abstract information goals because they reflect the invariant visual representations that enable the resemblance response, and must be broken down into global and local constituents according to the constraints of representation and implementation at each level of the visual hierarchy or their analogues in deep convolutional networks, where we can use a similar methodology to understand the identity contents represented in the hidden layers[33]. In norm-based coding[17,34], face-identity information is represented with reference to the average of a multidimensional face space. Monkey single-cell responses increase their firing rate with increasing distance of a face from this average (as happens with, for example, caricaturing[35],). As shown by Chang et al.[36], neurons selectively respond along a single axis of the face space, and not to other, orthogonal axes. An interesting direction of research is to determine whether our reduced diagnostic features, as defined by our abstract information goal (see also ref. [37]), provide a superior fit to the neural data than the full feature sets used in the axis model used by Chang et al.[36].

Though we modelled the mental representation of a face identity in an AAM, it is important to state that we do not assume that memory really represents faces in this way (that is, as demarcations to an average, separately for 3D shape and 2D texture). AAM is only a state-of-the-art, mathematical-modelling framework. We fully acknowledge there are many possible concrete implementations into a neural or a neurally inspired architecture that could deliver AAM-like performance without assuming an explicit AAM representation. What is clear is that whatever the implementation, in whichever architecture, the abstract information modelled under AAM framework will have to enable the performance characteristics demonstrated by our resemblance tasks.

For example, we would hypothesize that the diagnostic identity components in Fig. 3b are broken down, bottom to top, into the representational language of V1—that is, as representation in multiscale, multi-orientation Gabor-like, retinotopically mapped receptive fields[38,39] at intermediate levels of processing, as the type of local surface patches[40,41] that we reveal, and at the top level as the combinations of surface patches that enable identification and resemblance responses. Under a framework of top-down prediction[42,43], the abstract information goal of a familiar-face identity should trim, in a top-down manner, the fully mapped but redundant information on the retina into the task-relevant features that are transferred along the occipital to ventral–dorsal visual hierarchy[37]. Tracing the construction of such a reduced memory representation of face identity in the brain should enable an accurate and detailed modelling of the processing mechanism along the visual hierarchy (see also refs. [12,44–46]). What our work critically provides is an estimate of the end goal of the hierarchy (that is, the diagnostic component), which is also a prediction of what is important in the input. It is in this sense that mental representations guide task-specific information processing in the brain. Without knowing the mental representations, we do not even have an information needle to search in the fabled haystack of brain activity, let alone the ability to reconstruct the mechanisms that process its contents.

We modelled the critical mental representations that guide the processing of visual information of familiar-face identities. In several resemblance tasks that require usage of face identity, we demonstrated the efficacy of the contents we modelled. Our approach and results open new research avenues for the interplay between visual information, categorization tasks and their implementation as information-processing mechanisms in the brain.

Methods

Generative model of 3D face identity. We designed a generative model to objectively characterize and control 3D face-identity variance, using a database of 355 3D faces (acquired with a 4D face capture system, see '3D face database' in Supplementary Methods) that describes each face by its shape (with 3D coordinates for each one of 4,735 vertices) and its texture (with the RGB values of 800 × 600 pixels; see Supplementary Fig. 1a). It is critical to reiterate that the familiar faces were not part of the 3D face database.

To design the 3D GMF, we first applied a high-dimensional GLM, separately to 3D vertex coordinates and 2D pixel RGB values, to model and explain away variations in face shape and texture that arise from the non-identity categorical factors of sex, age, ethnicity and their interactions. The GLM therefore: (1) extracted as a non-identity face average the shape and texture face information explained by non-identity categorical factors; and (2) isolated the residual information that defines the 3D shape and 2D texture identity information of each face—that is, the identity residuals.

To further control identity information, we applied PCA to the identity residuals of the 355 faces, separately for shape and texture. The PCA represented shape residuals as a 355-dimensional vector in a 355-dimensional space of multivariate components, and a separate PCA represented the texture residuals as a 355 × 5 (spatial frequency bands)-dimensional matrix in a space of 355×5 multivariate components. Two sets of PCA coordinates therefore represented the objective shape and texture information of each identity in the principal components space of identity residuals.

Our 3D GMF is formally expressed as follows:

$$\text{Faces} = \text{Design matrix} \times \text{Coefficient matrix} + \text{Weights} \times \text{Principal components}$$

Where Faces is the vertex (or texture) matrix of 355 faces: for vertices, it is [355 × 14,205], where 14,205 = 4,735 vertices x 3 coordinates; for texture, it is [355 × 1,440,000], where 1,440,000 = 800 × 600 pixels x 3 RBG. Design matrix defined the non-identity categorical factors and their interactions ($n = 9$), that is, constant, age, gender, white Caucasian, eastern Asian, black African, gender × white Caucasian, gender × eastern Asian, gender × black African, for each of face ($n = 355$), and therefore is [355 × 9]. We estimated the linear effects of each non-identity factor and their interactions using the GLM; they are represented in the Coefficient matrix (that is, [9 × 14,205] for shape and [9 × 1,440,000] for texture). After the GLM fit, the [355 × 14,205] shape (or [355 × 140,000] texture) residuals are further explained using the PCA analysis, resulting in 355 components.

Supplementary Fig. 1b illustrates how the generative model controlled the non-identity and identity factors using the four familiar faces of our experiment. First, we scanned the four familiar faces of the experiment (second column). We fitted each into our 3D GMF to derive a ground truth face (the third column) with minimal distortions (shown in the first column).

The model generates new 3D faces by adding the identity residuals of four familiar faces to different non-identity GLM averages, to change their age, sex or ethnicity separately, or jointly sex and ethnicity. The outcomes are older, sex-swapped, ethnicity-swapped and sex-and-ethnicity-swapped versions of the same identity (the fourth to seventh columns). We used these generative properties to derive the stimuli of the generalization experiment.

Reverse-correlation experiment. *Participants.* We recruited 14 participants (all white Caucasians, seven females, age = 25.86 ± 2.26 yr (mean ± s.d.)) who were personally familiar with each familiar identity as work colleagues for at least six months. We assessed familiarity on a nine-point Likert scale, from not at all familiar (1) to highly familiar (9). Supplementary Table 1 reports the familiarity ratings for each identity and participant. We chose a sample size similar to those reported elsewhere[47–49]. All participants had normal or corrected-to-normal vision, without a self-reported history or symptoms of synaesthesia, and/or any psychological, psychiatric or neurological condition that affects face processing (for example, depression, autism spectrum disorder or prosopagnosia). They gave written informed consent and received £6 per hour for their participation. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval.

Familiar faces. We scanned four faces: 'Mary' and 'Stephany' (white Caucasian females, 36 and 38 yr of age, respectively), and 'John' and 'Peter' (white Caucasian males, 31 and 38 yr of age, respectively) who were familiar to all participants as work colleagues. As we will explain, we used these scanned faces to compare the objective and mentally represented identity information in each participant. Each of these four people gave informed consent for the use of their faces in published papers.

Random face identities. We reversed the flow of computation in the 3D GMF to synthesize new random identities while controlling their non-identity factors (see Fig. 1b, 'identity generation'; the reverse direction is indicated by the dashed line).

We proceeded in three steps: first, we fitted the familiar identity in the GLM to isolate its non-identity averages, independently for shape and texture. Second, we randomized identity information by creating random identity residuals—that is, we generated random coefficients (shape: 355; texture: 355 × 5) and multiplied them by the principal components of residual variance (shape: 355; texture: 355 × 5). Finally, we added the random identity residuals to the GLM averages to create a total of 10,800 random faces per familiar identity in the reverse-correlation experiment.

Procedure. Each experimental block started with a centrally presented frontal view of a randomly chosen familiar face (henceforth, the target). On each trial of the block, participants viewed six simultaneously presented randomly generated identities based on the target, displayed in a 2 × 3 array on a black background, with faces subtending an average of 9.5° by 6.4° of visual angle. We instructed participants to respond on one of six buttons to choose the face that most resembled the target. The six faces remained on the screen until there was a response. Another screen immediately followed, instructing participants to rank the similarity of their choice to the target, using a six-point Likert scale (1, not similar; 6, highly similar) with corresponding response buttons. Following the response, a new trial began. The experiment comprised 1,800 trials per target, divided into 90 blocks of 20 trials each, run over several days for a grand total of 7,200 trials that all validators accomplished in a random order. Throughout, participants sat in a dimly lit room and used a chin rest to maintain a 76 cm viewing distance. We ran the experiment using the Psychtoolbox for MATLAB R2012a. Data collection and following analysis were not performed blind to the target faces.

Analyses. *Linear regression model.* For each participant and target face, each trial produced two outcomes: one matrix of 4,735 × 3 vertex (and 800 × 600 RGB pixel) parameters corresponding to the shape (and texture) residuals of the chosen random face on this trial, and one corresponding integer that captures the similarity between the random identity parameters and the target. Across the 1,800 trials per target, we linearly regressed (using RobustFit, Matlab 2013b) the 3D residual vertices (separately for the *x*, *y* and *z* coordinates) and residual RGB pixels (separately for the R, G and B colour channels) with the corresponding similarity-rating values. These linear regressions produced a linear model with coefficients Beta_1 and Beta_2 vectors for each residual shape vertex coordinate and residual RGB texture pixel, for each familiar face and participant. Supplementary Fig. 2a illustrates the linear regression model for the 3D vertices of 'Mary'. Henceforth, we focus our analyses on the Beta_2 coefficients because they quantify how shape and texture identity residuals deviate from the GLM categorical average to represent the identity of each familiar face in the memory of each participant.

Reconstructing mental representations. Beta_2 coefficients can be amplified to control their relative presence in a newly synthesized 3D face. Supplementary Fig. 2b1 illustrates such amplification for one participant's Beta_2 coefficients of shape and texture of 'Mary'. Following the reverse-correlation experiment, we brought each participant back to fine-tune their Beta_2 coefficients for each familiar face, using the identical display and viewing distance parameters as in the reverse-correlation experiment (see Supplementary Fig. 2b2 and 'Fine-tuning Beta_2 coefficients' in Supplementary Methods).

Vertex contribution to mental representations. Vertices, whether in the ground truth face or in the participant's mental representation, can deviate inward or outward in 3D from the corresponding vertex in the common categorical average of their GLM fits (compare with Fig. 1b). Thus, we can compare the respective deviations of their 3D vertices in relation to the common GLM categorical average. To evaluate this relationship, we plotted the normalized deviation of ground truth vertices from most inward (−1) to most outward (+1) on the *x* axis of a 2D scatter plot; we also reported the normalized deviation of corresponding vertex of the mental representation on the *y* axis (as shown Fig. 2a). If ground truth and mental representations were identical, their vertex-by-vertex deviations from the GLM categorical average (that is, Euclidean distance) would be identical and would form the veridical diagonal straight white line provided as a reference in the scatter plot of Fig. 2a.

Using this veridical line as a reference, for each participant and familiar-face representation, we proceeded in three steps to classify each vertex as either 'faithful' or 'not faithful', and to test whether the vertices in mental representations deviated from the categorical average more than would be expected to occur by chance.

Step 1: we constructed a permutation distribution by iterating our regression analysis 1,000 times with random permutations of the choice response across the 1,800 trials. To control for multiple comparisons, we selected maximum (versus minimum) Beta_2 coefficients across all shape vertices (and texture pixels), separately for the *x*, *y* and *z* coordinates (RGB colour channels) from each iteration. We used the resulting distribution of maxima (and minima) to compute the 95% confidence interval of chance-level upper (and lower) Beta_2 value and classified each Beta_2 coefficient as significantly different from chance ($P < 0.05$, two-sided permutation test), or not. We consider the vertex (or pixel) as significant if the Beta_2 coefficient of any coordinate (or colour channel) was significant. There were very few significant pixels, with almost no consistency across participants (see Supplementary Fig. 3), so we excluded texture identity residuals from further analyses.

Step 2: we used the chance-fit Beta coefficients in step 1 and the Beta_2 amplification value derived in Reconstructing mental representation to compute the equation GLM + Beta_1 + Beta_2 × amplification value (compare with Supplementary Fig. 2b). As a result, we built a distribution of 1,000 chance-fit faces.

Step 3: to classify whether each significant 3D vertex in the mental representation of a participant is more similar to ground truth than we would expect by chance, we computed D_{chance}, the mean Euclidean distance between the 1,000 chance-fit faces and the veridical line, and D_{memory}, the distance between the same mental representation vertex and the veridical line. If $D_{\text{memory}} < D_{\text{chance}}$, this significant vertex is faithful because it is significantly closer to the veridical line than chance (and we plot it with blue to red colours in Fig. 2a); if $D_{\text{memory}} > D_{\text{chance}}$, the vertex is not faithful (and we plot it in white in Fig. 2a, together with the non-significant vertices).

To derive group results, we counted across participants the frequency of each faithful vertex and used a winner-take-all scheme to determine group-level consistency. For example, if 13 of 14 participants represented this particular vertex as faithful, we categorized it as such at the group level and reported the number of participants as a colour indicating 13 participants. If there was no majority for a vertex, we colour-coded it as white (see Fig. 2b).

Components of memory representation. The purpose of the following analysis was to find common diagnostic components (multivariate features) that emerged in the group-level memory representation of each face identity. To do so, we factorized with NNMF the total set of memory representations across familiar identities and observers.

For each participant, we recoded each vertex in the identity residuals of each familiar face as faithful (1), not faithful or not significant (0), resulting in a 4,735-dimensional binary vector. We pooled 56 such binary vectors (across 4 targets × 14 observers, that is, 56) to create a 4,735 × 56 (that is, vertex × model) binary matrix to which we applied NNMF to derive 8 multivariate components that captured the main features that faithfully represent familiar faces in memory across participants (see 'Non-negative matrix factorization' in Supplementary Methods). The heat map in Fig. 3a shows each NNMF component.

To determine the loading (that is, the contribution) of each NNMF component in the group-level mental representation of each familiar-face identity, we computed the median loading of this component on the 14 binary vectors representing this identity in the 14 observers. We applied a 0.1 loading threshold (>73 percentile of all 8 components × 4 identities median loadings) to ascribe a given component to a familiar-face representation. The box plot in Fig. 3a represents the loading of each NNMF component at the group-level representation, with coloured boxes showing at least two above-threshold NNMF components represent each familiar identity.

We then constructed the diagnostic component of a familiar-identity representation as follows: for each vertex, we extracted the maximum loading value across the NNMF components representing it, and normalized the values to the maximum loading across all vertices. This produced a 4,735-dimensional vector V_d, which weighs the respective contribution of each 3D vertex to the faithful representation of this familiar identity that we call the 'diagnostic component'. The heat maps in the left column of Fig. 3b represent the diagnostic component of each familiar identity. Supplementary Fig. 4 shows the high accuracy of the features captured by the components.

Crucially for our validation experiment, we were then able to define a non-diagnostic component (V_{n}) as the complement of the diagnostic component: $V_{\text{n}} = 1 - V_{\text{d}}$. It is important to emphasize that we adjusted the total deviation magnitude of the diagnostic and non-diagnostic components from the categorical average—that is, by equating the total sum of their deviations. This ensures that diagnostic and non-diagnostic components are both equidistant from the average face in the objective face space. The right column of Fig. 3b shows the non-diagnostic component of each familiar-identity representation.

Generalization experiments. *Validators.* We recruited 12 further participants (seven white Caucasian and one East Asian females; five white Caucasian males, aged 28.25 ± 4.11 yr (mean ± s.d.)), using the same procedure and criteria and those presiding for the selection of participants. Supplementary Table 2 reports the familiarity ratings for each identity and validator. All validators had normal or corrected-to-normal vision, without a self-reported history or symptoms of synaesthesia, and/or any psychological, psychiatric or neurological condition that affects face processing (for example, depression, autism spectrum disorder or prosopagnosia). They gave written informed consent and received £6 per hour for their participation. The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval.

Stimuli. For each familiar identity, we synthesized new 3D faces that comprised graded levels of either the diagnostic or the non-diagnostic shape components as explained in the section Components of memory representation above. Specifically, we used the normalized diagnostic component V_{d} and its non-diagnostic complement V_n to synthesize morphed faces with shape information of each target identity as follows:

$$\text{Diagnostic faces} = \text{Ground truth} \times V_d \times \alpha + \text{Categorical average } (1 - V_d \times \alpha)$$

$$\text{Non} - \text{diagnostic faces} = \text{Ground truth} \times V_n \times \alpha + \text{Categorical average } (1 - V_n \times \alpha)$$

with amplification factor $\alpha = 0.33, 0.67, 1, 1.33$ and 1.67, to control the relative intensity of diagnostic and non-diagnostic shape changes. We rendered all these morphed shapes with the same average texture. The first rows of Supplementary Figs. 5–8 show the morphed faces for each familiar identity. We added as filler stimuli the grand average face (for both shape and texture) of the 355 database faces.

We also changed the viewpoint, age and sex of all of these synthesized faces. Specifically, we rotated them in depth by −30°, 0° and +30°, and using the 3D GMF, we set the age factor to 80 yr or swapped the sex factor, keeping all other factors constant (compare with Generative model of 3D face identity in Fig. 1b and Supplementary Fig. 1b).

Procedure. The experiment comprised three sessions (viewpoint, age and sex) that all validators accomplished in a random order, with one session per day. In the viewpoint session, validators ran 15 blocks of 41 trials (five repetitions of 123 stimuli). Each trial started with a centrally displayed fixation for 1 s, followed by a face on a black background for 500 ms. We instructed validators to name the face as 'Mary', 'Stephany', 'John' or 'Peter' or respond 'other' if they could not identify the face. Validators were required to respond as accurately and as quickly as possible. A 2 s fixation separated each trial. Validators could break between blocks. In the age and sex sessions, validators ran five blocks that repeated 44 trials. They were instructed to respond "old Mary", "old Stephany", "old John", "old Peter" or "other" in the age session, and "Mary's brother", "Stephany's brother", "John's sister", "Peter's sister" or "other" in the sex session. For each session, stimuli were randomized across all trials. Across the three sessions, we recorded participants' identification performance in three viewpoints, a change of age information and a change of sex information. Data collection and subsequent analysis were not performed blind to the conditions of the experiments.

Analyses. For each validator and generalization condition, we computed the per cent correct identification of diagnostic and non-diagnostic faces for each familiar face and at each level of feature intensity. To ensure that diagnostic and non-diagnostic faces produced the expected effect for each one of the four identities, we fitted a linear mixed-effects model (that is, fitlme, Matlab 2016b) to the data of each identity separately, using Wilkinson's formulae:

$$\begin{aligned}\text{Performance} \approx\ & 1 + \text{Face type} + \text{Task type} + \text{Amplification}\\ & + (\text{Face type} + \text{Task type} + \text{Amplification} - 1\\ & \quad |\text{Subject})\end{aligned}$$

The model had fixed factors of Face type (that is, diagnostic versus non-diagnostic), feature Amplification (that is, 0.33, 0.67, 1, 1.33 and 1.67) and generalization Task type (that is, three views plus an age change and a sex change) as explanatory variables and participants' response variability as random factor. From this model, we can infer whether or not the fixed factors generalized beyond the specific participant sample, separately for each identity.

We tested the specified fixed effect factor (that is, using ANOVA, Matlab 2016b), using the Satherwither approximation to compute the approximate degrees of freedom. We found for each identity a higher identification performance with diagnostic than with non-diagnostic faces (see Fig. 4b), and the performance increased with amplification (an effect of feature amplification). The generalization task effect was significant for 'Mary' and 'Stephany' and not for 'John' and 'Peter'. Supplementary Tables 3–6 report the full statistics of our fixed effects, for each identity.

To further test the prediction effect of Face type we built a null model that excludes this factor:

$$\begin{aligned}\text{Performance} \approx\ & 1 + \text{Task type} + \text{Amplification}\\ & + (\text{Task type} + \text{Amplification} - 1|\text{Subject})\end{aligned}$$

For each identity, we compared the original and null model with a likelihood ratio (LR). Performance was significantly better explained by the original model (with Face type) than the null model (without Face type). For 'Mary', LR statistic = 603.72.135, $P < 0.001$; for 'Stephany', LR statistic = 39.516, $P < 0.001$; for 'John', LR statistic = 205.67, $P < 0.001$; for 'Peter', LR statistic = 214.34, $P < 0.001$. See Supplementary Tables 3–6 for the full statistical analysis.

We also found a significant interaction effect between Face type and Amplification, by fitting a linear mixed-effect model with this interaction included as an effect factor (see 'Linear mixed-effect model of face type by amplification interaction' in Supplementary Methods and Supplementary Table 7).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data are available in Mendeley Data with the identifier https://doi.org/10.17632/nyt677xwfm.1[50].

Code availability

Analysis scripts are available in Mendeley Data with the identifier https://doi.org/10.17632/nyt677xwfm.1[50].

Received: 20 September 2018; Accepted: 2 May 2019;
Published online: 17 June 2019

References

1. Bar, M. The proactive brain: memory for predictions. *Phil. Trans. R. Soc. B* **364**, 1235–1243 (2009).
2. Bar, M. et al. Top-down facilitation of visual recognition. *Proc. Natl Acad. Sci. USA* **103**, 449–454 (2006).
3. Ullman, S., Assif, L., Fetaya, E. & Harari, D. Atoms of recognition in human and computer vision. *Proc. Natl Acad. Sci. USA* **113**, 2744–2749 (2016).
4. Harel, A., Kravitz, D. J. & Baker, C. I. Task context impacts visual object processing differentially across the cortex. *Proc. Natl Acad. Sci. USA* **111**, E962–E971 (2014).
5. O'Toole, A. J. in *The Oxford Handbook of Face Perception* (eds Rhodes, G. et al.) 15–30 (Oxford Univ. Press, 2011).
6. Tsao, D. Y. & Livingstone, M. S. Mechanisms of face perception. *Annu. Rev. Neurosci.* **31**, 411–437 (2008).
7. Rosch, E. & Mervis, C. B. Family resemblances—studies in internal structure of categories. *Cogn. Psychol.* **7**, 573–605 (1975).
8. Ahumada, A. & Lovell, J. Stimulus features in signal detection. *J. Acoust. Soc. Am.* **49**, 1751 (1971).
9. Yu, H., Garrod, O. G. B. & Schyns, P. G. Perception-driven facial expression synthesis. *Comput. Graph.* **36**, 152–162 (2012).
10. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
11. Lee, H. & Kuhl, B. A. Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *J. Neurosci.* **36**, 6069–6082 (2016).
12. Nestor, A., Plaut, D. C. & Behrmann, M. Feature-based face representations and image reconstruction from behavioral and neural data. *Proc. Natl Acad. Sci. USA* **113**, 416–421 (2016).
13. Chang, C. H., Nemrodov, D., Lee, A. C. H. & Nestor, A. Memory and perception-based facial image reconstruction. *Sci. Rep.* **7**, 6499 (2017).
14. Turk, M. & Pentland, A. Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86 (1991).
15. Cootes, T. F., Edwards, G. J. & Taylor, C. J. Active appearance models. *IEEE Trans. Pattern Anal.* **23**, 681–685 (2001).
16. Blanz, V. & Vetter, T. A morphable model for the synthesis of 3D faces. In *Proc. 26th Annual Conference on Computer Graphics and Interactive Techniques* 187–194 (ACM Press/Addison–Wesley, 1999).
17. Rhodes, G. & Jeffery, L. Adaptive norm-based coding of facial identity. *Vis. Res* **46**, 2977–2987 (2006).
18. O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q. & Chellappa, R. Face space representations in deep convolutional neural networks. *Trends Cogn. Sci.* **22**, 794–809 (2018).
19. Young, A. W. & Burton, A. M. Are we face experts? *Trends Cogn. Sci.* **22**, 100–110 (2018).
20. White, D., Phillips, P. J., Hahn, C. A., Hill, M. & O'Toole, A. J. Perceptual expertise in forensic facial image comparison. *Proc. Biol. Sci.* **282**, 20151292 (2015).
21. Eger, E., Schweinberger, S. R., Dolan, R. J. & Henson, R. N. Familiarity enhances invariance of face representations in human ventral visual cortex: fMRI evidence. *Neuroimage* **26**, 1128–1139 (2005).
22. Jenkins, R., White, D., Van Montfort, X. & Burton, A. M. Variability in photos of the same face. *Cognition* **121**, 313–323 (2011).
23. Gosselin, F. & Schyns, P. G. RAP: a new framework for visual categorization. *Trends Cogn. Sci.* **6**, 70–77 (2002).
24. Schyns, P. G. Diagnostic recognition: task constraints, object information, and their interactions. *Cognition* **67**, 147–179 (1998).
25. Palmeri, T. J., Wong, A. C. N. & Gauthier, I. Computational approaches to the development of perceptual expertise. *Trends Cogn. Sci.* **8**, 378–386 (2004).
26. Burton, A. M., Schweinberger, S. R., Jenkins, R. & Kaufmann, J. M. Arguments against a configural processing account of familiar face recognition. *Perspect. Psychol. Sci.* **10**, 482–496 (2015).
27. Erens, R. G., Kappers, A. M. & Koenderink, J. J. Perception of local shape from shading. *Percept. Psychophys.* **54**, 145–156 (1993).
28. Phong, B. T. Illumination for computer generated pictures. *Commun. ACM* **18**, 311–317 (1975).

29. Liu, Z. L. Viewpoint dependency in object representation and recognition. *Spat. Vis.* **9**, 491–521 (1996).
30. Schyns, P. G., Goldstone, R. L. & Thibaut, J. P. The development of features in object concepts. *Behav. Brain Sci.* **21**, 1–17 (1998); discussion 17–54.
31. Mangini, M. C. & Biederman, I. Making the ineffable explicit: estimating the information employed for face classifications. *Cogn. Sci.* **28**, 209–226 (2004).
32. Baxter, M. G. Involvement of medial temporal lobe structures in memory and perception. *Neuron* **61**, 667–677 (2009).
33. Xu, T. et al. Deeper interpretability of deep networks. Preprint at https://arxiv.org/abs/1811.07807 (2018).
34. Leopold, D. A., O'Toole, A. J., Vetter, T. & Blanz, V. Prototype-referenced shape encoding revealed by high-level aftereffects. *Nat. Neurosci.* **4**, 89–94 (2001).
35. Leopold, D. A., Bondar, I. V. & Giese, M. A. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572–575 (2006).
36. Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).
37. Zhan, J., Ince, R. A. A., van Rijsbergen, N. & Schyns, P. G. Dynamic construction of reduced representations in the brain for perceptual decision behavior. *Curr. Biol.* **29**, 319–326 e314 (2019).
38. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–357 (2008).
39. Smith, F. W. & Muckli, L. Nonstimulated early visual areas carry information about surrounding context. *Proc. Natl Acad. Sci. USA* **107**, 20099–20103 (2010).
40. Peirce, J. W. Understanding mid-level representations in visual processing. *J. Vis.* **15**, 5 (2015).
41. Kubilius, J., Wagemans, J. & Op de Beeck, H. P. A conceptual framework of computations in mid-level vision. *Front. Comput. Neurosci.* **8**, 158 (2014).
42. Friston, K. J. & Kiebel, S. Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B* **364**, 1211–1221 (2009).
43. Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
44. Gosselin, F. & Schyns, P. G. Superstitious perceptions reveal properties of internal representations. *Psychol. Sci.* **14**, 505–509 (2003).
45. Smith, M. L., Gosselin, F. & Schyns, P. G. Measuring internal representations from behavioral and brain data. *Curr. Biol.* **22**, 191–196 (2012).
46. Nestor, A., Plaut, D. C. & Behrmann, M. Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proc. Natl Acad. Sci. USA* **108**, 9998–10003 (2011).
47. Gobbini, M. I. et al. Prioritized detection of personally familiar faces. *PLoS ONE* **8**, e66620 (2013).
48. van Belle, G., Ramon, M., Lefevre, P. & Rossion, B. Fixation patterns during recognition of personally familiar and unfamiliar faces. *Front. Psychol.* **1**, 20 (2010).
49. Ramon, M., Vizioli, L., Liu-Shuang, J. & Rossion, B. Neural microgenesis of personally familiar face recognition. *Proc. Natl Acad. Sci. USA* **112**, E4835–E4844 (2015).
50. Zhan, J., Garrod, O. G., van Rijsbergen, N. & Schyns, P. Modelling face memory reveals task-generalizable representations. *Mendeley Data* https://doi.org/10.17632/nyt677xwfm.1 (2019).

Acknowledgements

P.G.S. received support from the Wellcome Trust (Senior Investigator Award, UK; 107802) and the Multidisciplinary University Research Initiative/Engineering and Physical Sciences Research Council (USA, UK; 172046-01). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

J.Z., N.v.R. and P.G.S. designed the research; O.G.B.G. and P.G.S. developed the GMF; J.Z. performed the research; J.Z. and N.v.R. analysed the data; and J.Z., N.v.R. and P.G.S. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-019-0625-3.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to P.G.S.

Peer review information: Primary Handling Editor: Marike Schiffer

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Philippe G. Schyns

Last updated by author(s): 2019/04/29

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We ran the experiment using the Psychtoolbox for MATLAB R2012a.
Data analysis	We analysed the data using MATLAB R2013b & R2016b.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Data relating to current study are dposited in Mendeley Data with identifier http://dx.doi.org/10.17632/nyt677xwfm.1

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

In the Reverse Correlation Experiment, we modelled the 3D representational contents of 4 familiar faces in 14 participants' memory, by reverse correlating identity information generated on each trial with judgments of its similarity to the participant’s familiar face memory. In the Generalization Experiment, with new participants, we demonstrated the validity of these contents using everyday perceptual tasks that generalize face identity and resemblance judgments to new viewpoints, age and sex with a new group of participants.

Research sample

In the reverse correlation experiment, we recruited 14 participants (all white Caucasians, 7 females, mean age = 25.86 years, SD = 2.26 years) who were personally familiar with each familiar identity as work colleagues for at least 6 months.

In the generalization experiments, we recruited 12 validators (7 white Caucasian and 1 East Asian females, 5 white Caucasian males, with mean age = 28.25 years and SD = 4.11 years, using the same procedure and criteria and those presiding for the selection of participants.

Sampling strategy

We chose a sample size similar to those reported in literatures (Gobbini et al., 2013; Ramon, Vizioli, Liu-Shuang, & Rossion, 2015; van Belle, Ramon, Lefevre, & Rossion, 2010).

Data collection

All experiments are computer-based and therefore involve a PC, monitor, keyboard, and chin rest. We controlled the stimulus presentation and participant responses using MATLAB 2012b. Participants were located in a quiet standard testing room with light controlled. No one was present besides the participants during the testing.

In the Reverse Correlation Experiment, we used the data-driven approach and researchers have no strong hypothesis. In the Generalization Experiment,we used the within-subject design, in which trials from different conditions were randomized across task and researchers were blind to the trial order. The standard Task Instructions were provided in written form and were provided for each participant.

Timing

Reverse Correlation Experiment: 14/04/2015 - 15/08/2016
Generalization Experiment: 17/04/2018 - 15/06/2018

Data exclusions

No data were excluded from analysis.

Non-participation

No participants dropped out.

Randomization

Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

14 Western Caucasians (7 females, mean age = 25.86 years, SD = 2.26 years) participated in the reverse correlation experiment (experiment 1). 11 Western Caucasians (7 females) and 1 East Asian (female) participated in the generalization experiment (experiment 2, mean age = 28.25 years and SD = 4.11). All participants had a normal or corrected-to-normal vision, without a self-reported history or symptoms of synaesthesia, and/or any psychological, psychiatric or neurological condition that affects face processing (e.g., depression, autism spectrum disorder or prosopagnosia).

Recruitment

To target the participants who is familiar with the four testing identities, we used four faces from the School of Psychology, University of Glasgow, and circulated our adverts by emailing people who study or work in this school.

Ethics oversight

The University of Glasgow College of Science and Engineering Ethics Committee provided ethical approval.

Note that full information on the approval of the study protocol must also be provided in the manuscript.