



Zhan, J., Garrod, O. G.B., van Rijsbergen, N. and Schyns, P. G. (2019) Modelling face memory reveals task-generalizable representations. *Nature Human Behaviour*, 3, pp. 817-826. (doi:[10.1038/s41562-019-0625-3](https://doi.org/10.1038/s41562-019-0625-3))

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/188963/>

Deposited on: 25 June 2019

Enlighten – Research publications by members of the University of Glasgow
<http://eprints.gla.ac.uk>

16 Current cognitive theories are cast in terms of information processing mechanisms
17 that use mental representations [1-4]. For example, people use their mental
18 representations to identify familiar faces under various conditions of pose,
19 illumination and ageing, or to draw resemblance between family members. Yet, the
20 actual information contents of these representations are rarely characterized, which
21 hinders knowledge of the mechanisms that use them. Here, we modelled the 3D
22 representational contents of 4 faces that were familiar to 14 participants as work
23 colleagues. The representational contents were created by reverse correlating
24 identity information generated on each trial with judgments of the face's similarity to
25 the individual participant's memory of this face. In a second study, testing new
26 participants, we demonstrated the validity of the modelled contents using everyday
27 face tasks that generalize identity judgments to new viewpoints, age and sex. Our
28 work highlights that such models of mental representations are critical to
29 understanding generalization behavior and its underlying information processing
30 mechanisms.

31 The cognitive mechanism of recognition is guided by mental representations that are
32 stored in memory [1-4]. Personal familiarity with faces (e.g. as family members,
33 friends or work colleagues) provides a compelling everyday illustration because the
34 information contents representing familiar faces in memory must be sufficiently
35 detailed to enable accurate recognition (i.e. identifying 'Mary' amongst other people)
36 and sufficiently versatile to enable recognition across diverse common tasks—e.g.
37 identifying Mary in different poses, at different ages or identifying her brother based
38 on family resemblance [5-7]. And yet, it remains a fundamental challenge to reverse
39 engineer the participant's memory to model and thereby understand the detailed
40 contents of their representations of familiar faces. This challenge is a cornerstone to
41 understand the brain mechanisms of face identification, because they process the
42 contents to predict the appearance of the familiar face of 'Mary' in the visual array
43 and to selectively extract its identity information to generalize behavior across
44 common tasks.

45 We studied how our own work colleagues recognize the faces of other
46 colleagues from memory. The work environment provides a naturally occurring and
47 common medium of social interactions for all participants, who had at a minimum six
48 months of exposure with the people whose faces the study tested. To model the 3D
49 face identity information stored in their memory, we developed a methodology based
50 on reverse correlation (see Figure 1A, and *Methods, Reverse Correlation Experiment*)
51 and a new Generative Model of 3D Face Identity (i.e. GMF, see Figure 1B, and
52 *Methods, Generative Model of Face Identity*), separately for 3D shape and 2D texture
53 information (see Supplementary Figure 1A for 3D face parameters).

54 On each experimental trial, our GMF synthesized a set of 6 new 3D faces
55 (see Random Faces in Figure 1A), each with a unique and randomly generated
56 identity. Critically, each face shared other categorical face information (i.e. sex, age
57 and ethnicity) with one of the four faces that were personally familiar to each one of
58 our 14 participants as work colleagues—e.g. the familiar target face of 'Mary'. To
59 achieve this, we used a General Linear Model (GLM) to decompose the familiar
60 target face into a categorical component (e.g., for 'Mary' the average of all white
61 females faces of 30 years of age) plus a residual component that defines the specific
62 identity of the familiar face (see *Identity Modelling* in Figure 1B). We then generated
63 new random identities by keeping the categorical component of the target constant
64 (e.g., white female, 30 years of age) and adding a random component of identity (see
65 *Identity Generation* in Figure 1B, and *Methods, Reverse Correlation Experiment,*
66 *Random Face Identities* for details). Participants saw these randomly generated
67 faces in full frontal view and selected the one that most resembled the familiar target
68 (e.g., 'Mary') and rated its similarity to the target on a 6-point Likert scale, ranging
69 from not at all ('1') to highly similar ('6'). To resolve the task, participants must
70 compare the randomly generated faces presented on each trial with their mental
71 representation of the familiar target in full frontal view. Therefore, each face selected
72 comprises a match to the participant's mental representation of the target, which is
73 estimated by the similarity rating of that face.

74 After many such trials, we used reverse correlation [8] to estimate the
75 information content of the mental representation of each target familiar face (N= 4,
76 see Supplementary Figure 1B) in each participant (N = 14, see *Methods, Reverse*
77 *Correlation Experiment*). Specifically, we build a statistical relationship between the
78 information content of the faces that the participant selected on each trial with their
79 corresponding similarity ratings. In a second stage, we tested with a new group of
80 participants (N = 12, i.e. the validators, see *Methods, Generalization Experiments*)
81 whether these modelled mental representations were sufficiently detailed to enable
82 identification of each target familiar face and sufficiently versatile to enable
83 resemblance judgments across diverse everyday tasks--i.e. generalization across
84 new viewpoints, age and siblings.

85 To reconstruct the information contents of mental representations, we used
86 linear regression to compute the single-trial relationship between <similarity ratings,
87 random face identity components> for each target familiar face and participant.
88 Specifically, we computed separate regressions between the similarity ratings and
89 each 3D shape vertex and each RGB texture pixel that comprise the face identity
90 components. We then used the resulting Beta coefficients to model the 3D shape
91 and texture identity components that characterize the participant's mental
92 representation of each familiar face in the GMF (see Supplementary Figure 2 and
93 *Methods, Analyses, Linear Regression Model and Reconstructing Mental*
94 *Representations*).

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

115 Our analyses reveal the specific vertices near the veridical line that faithfully
116 represent 'Mary' in the mind of this participant as colored dots reported on the scatter
117 and located on the y-axis faces in Figure 2A. These vertices indicate faithful
118 representations because they are significantly closer to the ground truth faces than a
119 null distribution of representations arising from chance ($p < 0.05$, two-sided, with a
120 null distribution that iterated 1,000 times the analyses using a random permutation of
121 the participant's choice responses on each iteration, see details in *Methods*,
122 *Analyses, Vertex Contribution to Mental Representation*). In contrast, white vertices
123 away from the veridical line did not faithfully represent the identity. We repeated the
124 analysis of represented contents for each participant ($N = 14$) and familiar face ($N =$
125 4). Figure 2B reports the collated group results, using the format of Figure 2A, where
126 colors now indicate N , i.e. the number of participants who faithfully represented that
127 identity in their mind with this particular 3D shape vertex. Figure 2B demonstrates
128 that mental representations comprised similar information contents across the 14
129 individual participants. Most (10/14) faithfully represented 'Mary's' thin nose, 'John's'
130 receding eyes and wider upper face (13/14), 'Peter's' prominent eyebrow and jawline
131 (13/14), 'Stephany's' protruding mouth (13/14).

132 Such convergence of represented contents across participants suggests that
133 the face representations could be multivariate (i.e. comprising contiguous surface
134 patches rather than isolated vertices). As a final step, we extracted the main
135 multivariate components of represented surface patches. To this end, we applied
136 across observers ($N = 14$) and familiar faces ($N = 4$) the Non-negative Matrix
137 Factorization (NNMF, [10]) to the faithfully represented 3D vertices (see *Methods*,
138 *Analyses, Components of Memory Representation*). Figure 3A shows the multivariate
139 components that faithfully represent four target identities and Figure 3B shows their
140 combinations for the diagnostic components of each target identity (e.g. for 'Mary,'
141 the red background heatmap; for 'Stephany,' the green one and so forth). Importantly,
142 these diagnostic components of familiar face identity have complementary
143 nondiagnostic components (i.e. the grey background heatmaps in Figure 3B), which
144 capture variable face surfaces that do not comprise the participants' mental
145 representations.

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

157 For this demonstration, we synthesized new diagnostic (vs. nondiagnostic)
158 faces that were parametrically controlled for the relative strength of the diagnostic
159 multivariate components of identity vs. their nondiagnostic complement (see Figure
160 4A and *Methods, Generalization Experiments, Stimuli*). It is important to emphasize
161 that both diagnostic and nondiagnostic faces are equally faithful representations of
162 the original ground truth. That is, their shape features are equidistant from the shared
163 categorical average. However, whereas the diagnostic components deviate from the
164 average with multivariate information extracted from the participants' mental
165 representations, the nondiagnostic components do not. We hypothesized that,
166 though equidistant from the categorical average, only the diagnostic components will
167 impact performance on the resemblance tasks. For all synthesized faces, we
168 changed their viewpoint (rotation of -30 deg, 0 deg and +30 deg in depth), age (to 80
169 years old), and sex (to opposite) using the generative model—see Supplementary
170 Figure 5 to 8 for each familiar target.

171 In three independent resemblance tasks – changes of viewpoint, age and sex
172 – we tested the identification performance of 12 validators on the diagnostic and
173 nondiagnostic faces using a 5 Alternative Force Choice task (i.e. responding one of
174 four familiar identities plus a 'don't know' response, see *Methods, Generalization*
175 *Experiments, Procedure*). In each task, for each identity we found a significantly
176 higher identification performance for diagnostic faces (see Figure 4B, red curves)
177 than for nondiagnostic faces (black curves)—i.e. a fixed effect of Face Type in a
178 mixed effects linear model. For 'Mary', $F(1, 12.76) = 315.49, p < 0.001$, estimated
179 slope = 0.297, 95% Confidence Intervals = [0.264, 0.33]; for 'Stephany', $F(1, 20.62)$
180 $= 25.068, p < 0.001$, estimated slope = 0.058, 95% Confidence Intervals = [0.035,
181 0.081]; for 'John', $F(1, 12) = 21.369, p < 0.001$, estimated slope = 0.143, 95%
182 Confidence Intervals = [0.083, 0.204]; for 'Peter', $F(1, 12.01) = 5.76, p = 0.034$,
183 estimated slope = 0.095, 95% Confidence Intervals = [0.017, 0.173] (see *Methods,*
184 *Generalization Experiments, Analyses* for the detailed specification and
185 Supplementary Table 3 to 6 for the full statistical analysis of the models). Thus, the
186 diagnostic contents of the mental representations we modelled do indeed contain the
187 information that can resolve identity and resemblance tasks.

188 Mental representations stored in memory are critical to guide the information
189 processing mechanisms of cognition. Here, with a methodology based on reverse
190 correlation and a new 3D face information generator (i.e. our 3D GMF), we modelled
191 the information contents of mental representations of 4 familiar faces in 14 individual
192 participants. We showed that the contents converged across participants on a set of
193 multivariate features (i.e. local and global surface patches) that faithfully represent
194 3D information that is objectively diagnostic of each familiar face. Critically, we
195 showed that validators could identify new faces generated with these diagnostic
196 representations across three resemblance tasks—i.e. changes of pose, age and
197 sex—but performed much worse with equally faithful, but nondiagnostic features.
198 Together, our results demonstrate that the modelled representational contents were

199 both sufficiently precise to enable face identification within task and versatile enough
200 to generalize usage of the identity contents to other resemblance tasks.

201 At this stage, it worth stepping away from the results and emphasize that it is
202 remarkable that the reverse correlation methodology works at all, let alone produce
203 robust generalization across resemblance tasks. In the experiment, we asked
204 observers to rate the resemblance between a remembered familiar face, and
205 randomly generated faces, that by construction are very unlike the target face (never
206 identical, and almost never very similar). And yet, our results show that the
207 representational contents we modelled following such a task were in fact part of the
208 contents that objectively (i.e. faithfully) support identity recognition. This raises a
209 number of important points that we now discuss.

210 There has been a recent surge of interest in modelling face representations
211 from human memory [11-13]. These studies used 2D face images and applied
212 dimensionality reduction (e.g. PCA [14] and multidimensional scaling) to formalize an
213 image-based face space, where each dimension is a 2D eigenface or classification
214 image – i.e. pixel-wised RGB (or L*A*B) values. To understand the contribution of
215 each 2D face space dimension to memory representations (including their neural
216 coding), researchers modelled the relationship between projected weights of the
217 original 2D face images on each dimension and participants' corresponding
218 behavioral [13] (and brain [11, 12]) responses.

219 These studies contributed important developments in face identification
220 research because they addressed the face identity contents that the brain uses to
221 guide face identification mechanisms. Our aim was to model the face identity
222 contents in the generative 3D space of faces (not the 2D space of their image
223 projections) and to use these models to generate identification information in
224 resemblance tasks that test the generalizability of identity information. It is important
225 to clarify that we modelled identity information in a face space that belongs to the
226 broad class of 3D morphable, Active Appearance Models of facial synthesis (AAMs,
227 [15, 16]). These models contain full 3D surface and 2D texture information about
228 faces and so with their better control superseded the former generation of 2D image-
229 based face spaces ([14, 17] [18]). To synthesize faces, we used our GMF to
230 decompose each face identity as a linear combination of components of 3D shape
231 and 2D texture added to a local average (that summarizes the categorical factor of
232 age, gender, ethnicity and their interactions, cf. Figure 1B). To model the mental
233 representations of faces, we estimated the identity components of shape and texture
234 from the memory of each observer. These components had generative capacity and
235 we used them to precisely control the magnitude of identity information in new faces
236 synthesized to demonstrate generalization across pose, age and sex. Thus, we used
237 the same AAM framework for stimulus synthesis, mental representation estimation
238 and generation of generalizable identities.

239 There is a well-known problem with using AAMs to model the psychology of
240 face recognition. Perceptual expertise and familiarity are thought to involve

241 representations of faces that enable the greater generalization performance that is
242 widely reported [19-22]. However, AAMs typically adopt a brute force approach to
243 identity representation: a veridical (i.e. totally faithful) deviation of each physical
244 shape vertex and texture pixel from an average. Thus, as AAMs overfit identity
245 information, they appear as a priori weak candidate models to represent perceptual
246 expertise with faces [18]. Our approach of studying the contents of mental
247 representations suggests a solution to this conundrum. We showed that each
248 observer faithfully represented only a proportion of the objective identity information
249 that defines a familiar face identity. Our key theoretical contribution to face space is
250 to formalize the subjective 3D diagnostic information as a reduced set of multivariate
251 face features that can be construed as dimensions of the observer's face space.
252 Observers develop these dimensions when they interact with the objective
253 information that represents a new face identity in the real world. We modelled the
254 objective information that is available to the observer for developing their face space
255 dimensions via learning as the veridical shape and texture information of the AAM
256 [18, 23, 24]. Key to demonstrating the psychological relevance of our psychological
257 3D face space dimensions is that they should comprise identity information
258 sufficiently detailed to enable accurate face identification and sufficiently versatile to
259 enable similarity judgments of identity in resemblance tasks. We demonstrated this
260 potential when validators identified faces synthesized with the diagnostic dimensions
261 in novel resemblance tasks. Thus, by introducing reduced faithful mental
262 representations of identity information in the objective representations of AAMs we
263 provide the means of modelling the subjective psychological dimensions of an
264 individual's face space.

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

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

300 A general question with reverse correlation tasks is whether the resulting
301 models represent a particular visual category (here, the visual identity of a face) or
302 the task from which the model was reconstructed [24, 29-31]. We contributed to this
303 debate by showing that the identity information reconstructed in one task had efficacy
304 in other tasks that involved identity. Importantly, the tasks were designed to test two
305 classes of factors: ambient and categorical. For example, we showed that the identity
306 component extracted in one ambient viewpoint (full face, 0 deg) could be used to
307 generalize identification of the same face under two new ambient viewpoints (-30 and
308 +30 deg of rotation in depth). We also showed that the identity component extracted
309 for identities (all < 40 years of age) generalized to older age (80 years). Furthermore,
310 we also showed that though extracted from a given sex, the identity component
311 would generalize to another sex, a kinship task. Hence, we found no dramatic
312 differences due to the effect of task of extraction of the identity component. Rather,
313 the extracted representational basis is useful for all tasks tested, whether using
314 ambient or categorical factors of face variance. This therefore suggests that we have
315 tapped into some essential information about familiar face representation. However,
316 we acknowledge that the generalizations we observe might still be a function of an
317 interaction between the nature of memory and the similarity task from which we
318 estimated the identity component. The component could have differed had the task
319 been more visual than memory based (e.g. identification of the same face under
320 different orientations, or a visual matching task) and we might not have derived an
321 identity component that enabled such effective generalization. In any case, the
322 memorized identity components that enable task generalization reflect an interaction
323 between memory and the input information available to represent this identity [24, 32].
324 Observers can compare this memory representation for that identity with a
325 representation of the visual input for successful identification.

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

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

352 For example, we would hypothesize that the diagnostic identity components
353 in Figure 3B are broken down, bottom to top, into the representational language of
354 V1—i.e. as representation in multi-scale, multi-orientation Gabor-like, retinotopically
355 mapped receptive fields [38, 39]; at intermediate levels of processing, as the sort of
356 local surface patches [40, 41] that we reveal, and at the top level as the combinations
357 of surface patches that enable identification and resemblance responses. Under a
358 framework of top-down prediction [42, 43], the abstract information goal of a familiar
359 face identity should trim, in a top-down manner, the fully-mapped but redundant
360 information on the retina into the task-relevant features that are transferred along the
361 occipital to ventral/dorsal visual hierarchy [37]. Tracing the construction of such a
362 reduced memory representation of face identity in the brain should enable an
363 accurate and detailed modelling of the processing mechanism along the visual
364 hierarchy (see also [12, 44-46]). What our work critically provides is an estimate of
365 the end goal of the hierarchy (i.e. the diagnostic component), which is also a
366 prediction of what is important in the input. It is in this sense that mental
367 representations guide task-specific information processing in the brain. Without
368 knowing mental representations, we do not have even have an information needle to

369 search in the fabled haystack of brain activity, let alone reconstruct the mechanisms
370 that process its contents.

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

377 **METHODS**

378 **Generative Model of 3D Face Identity (GMF).**

379 We designed a generative model to objectively characterize and control 3D face
380 identity variance, using a database of 355 3D faces (acquired with a 4D face capture
381 system, see *Supplementary Methods, 3D Face Database*) that describes each face
382 by its shape (with 3D coordinates for each one of 4,735 vertices) and its texture (with
383 the RGB values of 800*600 pixels, see Supplementary Figure 1A). It is critical to
384 reiterate that the familiar faces were not part of the 3D face database.

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

393 To further control identity information, we applied Principal Components
394 Analysis (PCA) to the identity residuals of the 355 faces, separately for shape and
395 texture. The PCA represented shape residuals as a 355-dimensional vector in a 355-
396 dimensional space of multivariate components, and a separate PCA represented the
397 texture residuals as a 355*5 (spatial frequency bands)-dimensional matrix in a space
398 of 355*5 multivariate components. Two sets of PCA coordinates therefore
399 represented the objective shape and texture information of each identity in the
400 principal components space of identity residuals.

401 Our 3D GMF is formally expressed as follows:

$$Faces = Design\ Matrix \times Coefficient\ Matrix + weights \times PCs$$

402 Where *Faces* is the vertex (or texture) matrix of 355 faces: for vertices, it is
403 [355 x 14,205] where 14,205 = 4,735 vertices x 3 coordinates; for texture, it is [355 x
404 1,440,000] where 1,440,000 = 800 x 600 pixels x 3 RGB. *Design Matrix* defined the
405 non-identity categorical factors and their interactions (N = 9), i.e. constant, age,
406 gender, white Caucasian (WC), eastern Asian (EA), black African (BA), gender x WC,
407 gender x EA, gender x BA, for each of face (N = 355), and therefore is [355 x 9]. We
408 estimated the linear effects of each non-identity factor and their interactions using the
409 GLM which are represented in the *Coefficient Matrix* (i.e. [9 x 14,205] for shape and
410 [9 x 1,440,000] for texture). After the GLM fit, the [355 x 14,205] shape (or [355 x
411 140,000] texture) residuals are further explained using the PCA analysis, resulting
412 355 components.

413 Furthermore, Supplementary Figure 1B illustrates how the generative model
414 controlled the non-identity and identity factors using the 4 familiar faces of our

415 experiment. First, we scanned the four familiar faces of the experiment (2nd column).
416 We fitted each into our 3D GMF to derive a ground truth face (the 3rd column), with
417 minimal distortions (shown in the 1st column).

418 The model generates new 3D faces by adding the identity residuals of four
419 familiar faces to different non-identity GLM averages, to change their age, sex or
420 ethnicity separately, or jointly sex and ethnicity. The outcomes are older, sex
421 swapped, ethnicity swapped and sex and ethnicity swapped versions of the same
422 identity (the 4th to 7th column). We used these generative properties to derive the
423 stimuli of the generalization experiment.

424 **Reverse Correlation Experiment**

425 **Participants.** We recruited 14 participants (all white Caucasians, 7 females,
426 mean age = 25.86 years, SD = 2.26 years) who were personally familiar with each
427 familiar identity as work colleagues for at least 6 months. We assessed familiarity on
428 a 9-point Likert scale, from not at all familiar '1' to highly familiar '9'. Supplementary
429 Table 1 reports the familiarity ratings for each identity and participant. We chose a
430 sample size similar to those reported elsewhere [47-49]. All participants had normal
431 or corrected-to-normal vision, without a self-reported history or symptoms of
432 synaesthesia, and/or any psychological, psychiatric or neurological condition that
433 affects face processing (e.g., depression, autism spectrum disorder or
434 prosopagnosia). They gave written informed consent and received £6 per hour for
435 their participation. The University of Glasgow College of Science and Engineering
436 Ethics Committee provided ethical approval.

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

444 **Random Face Identities.** We reversed the flow of computation in the 3D
445 GMF to synthesize new random identities while controlling their non-identity factors
446 (see Figure 1B *Identity Generation*, the reverse direction is indicated by the dashed
447 line). We proceeded in three steps: First, we fitted the familiar identity in the GLM to
448 isolate its non-identity averages, independently for shape and texture. Second, we
449 randomized identity information by creating random identity residuals—i.e. we
450 generated random coefficients (shape: 355; texture: 355*5) and multiplied them by
451 the principal components of residual variance (shape: 355; texture: 355*5). Finally,
452 we added the random identity residuals to the GLM averages to create a total of
453 10,800 random faces per familiar identity in the reverse correlation experiment.

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

470 **Analyses**

471 **Linear Regression Model.** For each participant and target face, each trial
472 produced two outcomes: one matrix of 4,735*3 vertex (and 800*600 RGB pixel)
473 parameters corresponding to the shape (and texture) residuals of the chosen random
474 face on this trial, and one corresponding integer that captures the similarity between
475 the random identity parameters and the target. Across the 1,800 trials per target, we
476 linearly regressed (i.e. RobustFit, Matlab 2013b) the 3D residual vertices (separately
477 for the X, Y and Z coordinates) and residual RGB pixels (separately for R, G and B
478 color channel) with the corresponding similarity rating values. These linear
479 regressions produced a linear model with coefficients Beta_1 and Beta_2 vectors for
480 each residual shape vertex coordinate and residual RGB texture pixel, for each
481 familiar face and participant. Supplementary Figure 2A illustrates the linear
482 regression model for the 3D vertices of 'Mary.' Henceforth, we focus our analyses on
483 the Beta_2 coefficients because they quantify how shape and texture identity
484 residuals deviate from the GLM categorical average to represent the identity of each
485 familiar face in the memory of each participant.

486 **Reconstructing Mental Representations.** Beta_2 coefficients can be
487 amplified to control their relative presence in a newly synthesized 3D face.
488 Supplementary Figure 2B1 illustrates such amplification for one participant's Beta_2
489 coefficients of shape and texture of 'Mary.' Following the reverse correlation
490 experiment, we brought each participant back to fine-tune their Beta_2 coefficients
491 for each familiar face, using the identical display and viewing distance parameters as
492 in the reverse correlation experiment (see Supplementary Figure 2B2 and
493 *Supplementary Methods, Fine-tuning Beta_2 Coefficients*).

494 **Vertex Contribution to Mental Representations.** Vertices, whether in the
495 ground truth face or in the participant's mental representation can deviate inward or

496 outward in 3D from the corresponding vertex in the common categorical average of
497 their GLM fits (cf. Figure 1B). Thus, we can compare the respective deviations of
498 their 3D vertices in relation to the common GLM categorical average. To evaluate
499 this relationship, we plotted the normalized deviation of ground truth vertices from
500 most Inward (-1) to most Outward (+1) on the X-axis of a 2D scatter plot; we also
501 reported the normalized deviation of corresponding vertex of the mental
502 representation on the Y-axis (as shown Figure 2A). If ground truth and mental
503 representations were identical, their vertex-by-vertex deviations from the GLM
504 categorical average (i.e. Euclidean distance) would be identical and would form the
505 veridical diagonal straight white line provided as a reference in the scatter plot of
506 Figure 2A.

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

511 Step 1: We constructed a permutation distribution by iterating our regression
512 analysis 1,000 times with random permutations of the choice response across the
513 1,800 trials. To control for multiple comparisons, we selected maximum (vs. minimum)
514 Beta_2 coefficients across all shape vertices (and texture pixels), separately for the X,
515 Y and Z coordinates (RGB color channels) from each iteration. We used the resulting
516 distribution of maxima (and minima) to compute the 95% confidence interval of
517 chance-level upper (and lower) Beta_2 value and classified each Beta_2 coefficient
518 as significantly different from chance ($p < 0.05$, two-sided), or not. We consider the
519 vertex (or pixel) as significant if the Beta_2 coefficient of any coordinate (or color
520 channel) was significant. There were very few significant pixels, with almost no
521 consistency across participants (see Supplementary Figure 3), so we excluded
522 texture identity residuals from further analyses.

523 Step 2: We used the chance-fit Beta coefficients in Step 1 and the Beta_2
524 amplification value derived in **Reconstructing Mental Representation** to compute
525 the equation $GLM + \beta_1 + \beta_2 * amplification\ value$ (cf. Supplementary Figure 2B). As
526 a result, we built a distribution of 1,000 chance fit faces.

527 Step 3: To classify whether each significant 3D vertex in the mental
528 representation of a participant is more similar to ground truth than we would expect
529 by chance, we computed D_{chance} , the mean Euclidean distance between the 1,000
530 chance fit faces and the veridical line, and D_{memory} , the distance between the same
531 mental representation vertex and the veridical line. If $D_{memory} < D_{chance}$, this
532 significant vertex is 'faithful' because it is significantly closer to the veridical line than
533 chance (and we plot it with blue to red colors in Figure 2A); if $D_{memory} > D_{chance}$, the
534 vertex is not faithful (and we plot it in white in Figure 2A, together with the
535 nonsignificant vertices).

536 To derive group results, we counted across participants the frequency of each
537 faithful vertex and used a Winner-Take-All scheme to determine group-level
538 consistency. For example, if 13/14 participants represented this particular vertex as
539 'faithful,' we categorized it as such at the group level and reported the number of
540 participants as a color indicating 13 participants. If there was no majority for a vertex,
541 we color-coded it as white (see Figure 2B).

542 **Components of Memory Representation.** The purpose of the following
543 analysis was to find common diagnostic components (multivariate features) that
544 emerged in the group-level memory representation of each face identity. To do so,
545 we factorized with Non-negative Matrix Factorization (NNMF) the total set of memory
546 representations across familiar identities and observers.

547 For each participant, we recoded each vertex in the identity residuals of each
548 familiar face as 'faithful' = 1, 'not faithful' or not significant = 0, resulting in a 4735-d
549 binary vector. We pooled 56 such binary vectors (across 4 targets x 14 observers =
550 56) to create a 4735 by 56 (i.e. vertex-by-model) binary matrix to which we applied
551 NNMF to derive 8 multivariate components that captured the main features that
552 faithfully represent familiar faces in memory across participants (see *Supplementary*
553 *Methods, Non-negative Matrix Factorization*). Heatmap in Figure 3A shows each
554 NNMF component.

555 To determine the loading (i.e. the contribution) of each NNMF component in
556 the group-level mental representation of each familiar face identity, we computed the
557 median loading of this component on the 14 binary vectors representing this identity
558 in the 14 observers. We applied a 0.1 loading threshold (> 73 percentile of all 8
559 components × 4 identities median loadings) to ascribe a given component to a
560 familiar face representation. The boxplot in Figure 3A represents the loading of each
561 NNMF component at the group-level representation, with colored boxes showing at
562 least 2 above-threshold NNMF components represent each familiar identity.

563 We then constructed the diagnostic component of a familiar identity
564 representation as follows: for each vertex we extracted the maximum loading value
565 across the NNMF components representing it, and normalized the values to the
566 maximum loading across all vertices. This produced a 4735-d vector V_d that weighs
567 the respective contribution of each 3D vertex to the faithful representation of this
568 familiar identity that we call the "diagnostic component." The heat maps in the left
569 column of Figure 3B represent the diagnostic component of each familiar identity.
570 Supplementary Figure 4 shows the high accuracy of the features captured by the
571 components.

572 Crucially for our validation experiment, we were then able to define a
573 nondiagnostic component as the complement of the diagnostic component $V_n = 1 -$
574 V_d . It is important to emphasize that we adjusted the total deviation magnitude of the
575 diagnostic and nondiagnostic components from the categorical average—i.e. by
576 equating the total sum of their deviations. This ensures that diagnostic and

577 nondiagnostic components are both equidistant from the average face in the
578 objective face space. The right column of Figure 3B shows the nondiagnostic
579 component of each familiar identity representation.

580 **Generalization Experiments**

581 **Validators.** We recruited 12 further participants (7 white Caucasian and 1
582 East Asian females, 5 white Caucasian males, with mean age = 28.25 years and SD
583 = 4.11 years), using the same procedure and criteria and those presiding for the
584 selection of participants. Supplementary Table 2 reports the familiarity ratings for
585 each identity and validator. All validators had normal or corrected-to-normal vision,
586 without a self-reported history or symptoms of synaesthesia, and/or any
587 psychological, psychiatric or neurological condition that affects face processing (e.g.,
588 depression, autism spectrum disorder or prosopagnosia). They gave written informed
589 consent and received £6 per hour for their participation. The University of Glasgow
590 College of Science and Engineering Ethics Committee provided ethical approval.

591 **Stimuli.** For each familiar identity, we synthesized new 3D faces that
592 comprised graded levels of either the diagnostic or the nondiagnostic shape
593 components as explained in the section **Components of Memory Representation**
594 above. Specifically, we used the normalized diagnostic component V_d and its
595 nondiagnostic complement V_n to synthesize morphed faces with shape information of
596 each target identity as follows:

$$\text{Diagnostic Faces} = \text{Ground Truth} \times V_d \times \alpha + \text{Categorical Average} (1 - V_d \times \alpha)$$

$$\text{Nondiagnostic Faces} = \text{Ground Truth} \times V_n \times \alpha + \text{Categorical Average} (1 - V_n \times \alpha)$$

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

603 We also changed the viewpoint, age and sex of all of these synthesized faces.
604 Specifically, we rotated them in depth by -30 deg, 0 deg and +30 deg and using the
605 3D GMF, we set the age factor to 80 years/swapped the sex factor, keeping all other
606 factors constant (cf. *Generative Model of 3D Face Identity* in Figure 1B and
607 Supplementary Figure 1B).

608 **Procedure.** The experiment comprised 3 sessions (viewpoint, age and sex)
609 that all validators accomplished in a random order, with one session per day. In the
610 Viewpoint session, validators ran 15 blocks of 41 trials (5 repetitions of 123 stimuli).
611 Each trial started with a centrally displayed fixation for 1s, followed by a face on a
612 black background for 500ms. We instructed validators to name the face as 'Mary,'
613 'Stephany,' 'John' or 'Peter,' or respond 'other' if they could not identify the face.

614 Validators were required to respond as accurately and as quickly as possible. A 2s
615 fixation separated each trial. Validators could break between blocks. In the Age and
616 Sex sessions, validators ran 5 blocks that repeated 44 trials. They were instructed to
617 respond “Old Mary,” “Old Stephany,” “Old John,” “Old Peter” or “Other” in the age
618 session, and “Mary’s brother,” “Stephany’s brother,” “John’s sister,” “Peter’s sister” or
619 “Other” in the sex session. For each session, stimuli are randomized across all trials.
620 Across the 3 sessions, we recorded participants’ identification performance in 3
621 viewpoints, a change of age information and a change of sex information. Data
622 collection and following analysis were not performed blind to the conditions of the
623 experiments.

624 **Analyses.** For each validator and generalization condition, we computed the
625 percent correct identification of diagnostic and nondiagnostic faces for each familiar
626 face and at each level of feature intensity. To ensure that diagnostic and
627 nondiagnostic faces produced the expected effect for each one of the four identities,
628 we fitted a linear mixed effects model (i.e. fitlme, Matlab 2016b) to the data of each
629 identity separately, using Wilkinson’s formulae:

$$\text{Performance} \sim 1 + \text{Face Type} + \text{Task Type} + \text{Amplification} \\ + (\text{Face Type} + \text{Task Type} + \text{Amplification} - 1 | \text{Subject})$$

630 The model had fixed factors of Face Type (i.e. diagnostic vs. nondiagnostic),
631 Feature Amplification (i.e. 0.33, 0.67, 1, 1.33, 1.67) and Generalization Task (i.e. 3
632 views plus an age change and a sex change) as explanatory variables and
633 participants’ response variability as random factor. From this model, we can infer
634 whether or not the fixed factors generalized beyond the specific participant sample,
635 separately for each identity.

636 We tested the specified fixed effect factor (i.e. using ANOVA, Matlab 2016b),
637 using the Satherwith approximation to compute the approximate degrees of
638 freedom. We found for each identity a higher identification performance with
639 diagnostic than nondiagnostic faces (see Figure 4B), and the performance increased
640 with amplification (an effect of Feature Amplification). The Generalization Task effect
641 was significant for ‘Mary’ and ‘Stephany’ and not for ‘John’ and ‘Peter’.
642 Supplementary Table 3 to 6 report the full statistics of our fixed effects, for each
643 identity.

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

$$\text{Performance} \sim 1 + \text{Task Type} + \text{Amplification} + (\text{Task Type} + \text{Amplification} - 1 | \text{Subject})$$

646 For each identity, we compared the original and null model with a likelihood
647 ratio (i.e. LR). Performance was significantly better explained by the original model
648 (with Face Type) than the null model (without Face Type). For ‘Mary’, LR statistic =
649 603.72.135, $p < 0.001$; for ‘Stephany’, LR statistic = 39.516, $p < 0.001$; for ‘John’, LR

650 statistic = 205.67, $p < 0.001$; for 'Peter', LR statistic = 214.34, $p < 0.001$. See
651 Supplementary Table 3 to 6 for the full statistical analysis.

652 We also found a significant interaction effect between Face Type and
653 Amplification, by fitting a linear mixed effect model with this interaction included as an
654 effect factor (see Supplementary Methods, Linear Mixed Effect Model of Face Type
655 by Amplification Interaction, and Supplementary Table 7).

656 **Data Availability.** Data is available in Mendeley Data with identifier
657 <http://dx.doi.org/10.17632/nyt677xwfm.1> [50].

658 **Code Availability.** Analysis scripts are available in Mendeley Data with identifier
659 <http://dx.doi.org/10.17632/nyt677xwfm.1> [50].

660

661 REFERENCES

- 662 1. Bar, M. (2009). The proactive brain: memory for predictions. *Philos T R Soc B* *364*,
663 1235-1243.
- 664 2. Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J., Schmid, A.M., Dale, A.M.,
665 Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., et al. (2006). Top-down
666 facilitation of visual recognition. *Proc Natl Acad Sci U S A* *103*, 449-454.
- 667 3. Ullman, S., Assif, L., Fetaya, E., and Harari, D. (2016). Atoms of recognition in human
668 and computer vision. *P Natl Acad Sci USA* *113*, 2744-2749.
- 669 4. Harel, A., Kravitz, D.J., and Baker, C.I. (2014). Task context impacts visual object
670 processing differentially across the cortex. *Proc Natl Acad Sci U S A* *111*, E962-971.
- 671 5. O'Toole, A.J. (2011). Cognitive and Computational Approaches to Face Recognition In
672 *The Oxford Handbook of Face Perception*, G. Rhodes, A. Calder, M. Johnson and J.V.
673 Haxby, eds., pp. 15 -30.
- 674 6. Tsao, D.Y., and Livingstone, M.S. (2008). Mechanisms of face perception. *Annu Rev*
675 *Neurosci* *31*, 411-437.
- 676 7. Rosch, E., and Mervis, C.B. (1975). Family Resemblances - Studies in Internal
677 Structure of Categories. *Cognitive Psychol* *7*, 573-605.
- 678 8. Ahumada, A., and Lovell, J. (1971). Stimulus Features in Signal Detection. *J Acoust*
679 *Soc Am* *49*, 1751-&.
- 680 9. Yu, H., Garrod, O.G.B., and Schyns, P.G. (2012). Perception-driven facial expression
681 synthesis. *Comput Graph-Uk* *36*, 152-162.
- 682 10. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative
683 matrix factorization. *Nature* *401*, 788-791.
- 684 11. Lee, H., and Kuhl, B.A. (2016). Reconstructing Perceived and Retrieved Faces from
685 Activity Patterns in Lateral Parietal Cortex. *J Neurosci* *36*, 6069-6082.
- 686 12. Nestor, A., Plaut, D.C., and Behrmann, M. (2016). Feature-based face
687 representations and image reconstruction from behavioral and neural data. *Proc*
688 *Natl Acad Sci U S A* *113*, 416-421.

- 689 13. Chang, C.H., Nemrodov, D., Lee, A.C.H., and Nestor, A. (2017). Memory and
690 Perception-based Facial Image Reconstruction. *Sci Rep-Uk* 7.
- 691 14. Turk, M., and Pentland, A. (1991). Eigenfaces for recognition. *J Cogn Neurosci* 3, 71-
692 86.
- 693 15. Cootes, T.F., Edwards, G.J., and Taylor, C.J. (2001). Active appearance models. *IEEE T*
694 *Pattern Anal* 23, 681-685.
- 695 16. Blanz, V., and Vetter, T. (1999). A morphable model for the synthesis of 3D faces.
696 *Comp Graph*, 187-194.
- 697 17. Rhodes, G., and Jeffery, L. (2006). Adaptive norm-based coding of facial identity.
698 *Vision Res* 46, 2977-2987.
- 699 18. O'Toole, A.J., Castillo, C.D., Parde, C.J., Hill, M.Q., and Chellappa, R. (2018). Face
700 Space Representations in Deep Convolutional Neural Networks. *Trends Cogn Sci* 22,
701 794 - 809.
- 702 19. Young, A.W., and Burton, A.M. (2018). Are We Face Experts? *Trends in Cognitive*
703 *Sciences* 22, 100-110.
- 704 20. White, D., Phillips, P.J., Hahn, C.A., Hill, M., and O'Toole, A.J. (2015). Perceptual
705 expertise in forensic facial image comparison. *Proc Biol Sci* 282.
- 706 21. Eger, E., Schweinberger, S.R., Dolan, R.J., and Henson, R.N. (2005). Familiarity
707 enhances invariance of face representations in human ventral visual cortex: fMRI
708 evidence. *Neuroimage* 26, 1128-1139.
- 709 22. Jenkins, R., White, D., Van Montfort, X., and Burton, A.M. (2011). Variability in
710 photos of the same face. *Cognition* 121, 313-323.
- 711 23. Gosselin, F., and Schyns, P.G. (2002). RAP: a new framework for visual categorization.
712 *Trends Cogn Sci* 6, 70-77.
- 713 24. Schyns, P.G. (1998). Diagnostic recognition: task constraints, object information, and
714 their interactions. *Cognition* 67, 147-179.
- 715 25. Palmeri, T.J., Wong, A.C.N., and Gauthier, I. (2004). Computational approaches to
716 the development of perceptual expertise. *Trends in Cognitive Sciences* 8, 378-386.
- 717 26. Burton, A.M., Schweinberger, S.R., Jenkins, R., and Kaufmann, J.M. (2015).
718 Arguments Against a Configural Processing Account of Familiar Face Recognition.
719 *Perspect Psychol Sci* 10, 482-496.
- 720 27. Erens, R.G., Kappers, A.M., and Koenderink, J.J. (1993). Perception of local shape
721 from shading. *Percept Psychophys* 54, 145-156.
- 722 28. Phong, B.T. (1975). Illumination for Computer Generated Pictures. *Commun Acm* 18,
723 311-317.
- 724 29. Liu, Z.L. (1996). Viewpoint dependency in object representation and recognition.
725 *Spatial Vision* 9, 491-521.
- 726 30. Schyns, P.G., Goldstone, R.L., and Thibaut, J.P. (1998). The development of features
727 in object concepts. *Behav Brain Sci* 21, 1-17; discussion 17-54.
- 728 31. Mangini, M.C., and Biederman, I. (2004). Making the ineffable explicit: estimating
729 the information employed for face classifications. *Cognitive Sci* 28, 209-226.
- 730 32. Baxter, M.G. (2009). Involvement of medial temporal lobe structures in memory and
731 perception. *Neuron* 61, 667-677.

- 732 33. Xu, T., Zhan, J., Garrod, O.G.B., Torr, P.H.S., Zhu, S.C., Ince, R.A., and Schyns, P.G.
733 (2018). Deeper Interpretability of Deep Networks. ArXiv.
- 734 34. Leopold, D.A., O'Toole, A.J., Vetter, T., and Blanz, V. (2001). Prototype-referenced
735 shape encoding revealed by high-level aftereffects. *Nat Neurosci* 4, 89-94.
- 736 35. Leopold, D.A., Bondar, I.V., and Giese, M.A. (2006). Norm-based face encoding by
737 single neurons in the monkey inferotemporal cortex. *Nature* 442, 572-575.
- 738 36. Chang, L., and Tsao, D.Y. (2017). The Code for Facial Identity in the Primate Brain.
739 *Cell* 169, 1013-1028 e1014.
- 740 37. Zhan, J., Ince, R.A.A., van Rijsbergen, N., and Schyns, P.G. (2019). Dynamic
741 Construction of Reduced Representations in the Brain for Perceptual Decision
742 Behavior. *Curr Biol* 29, 319-326 e314.
- 743 38. Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural
744 images from human brain activity. *Nature* 452, 352-U357.
- 745 39. Smith, F.W., and Muckli, L. (2010). Nonstimulated early visual areas carry
746 information about surrounding context. *P Natl Acad Sci USA* 107, 20099-20103.
- 747 40. Peirce, J.W. (2015). Understanding mid-level representations in visual processing. *J*
748 *Vis* 15, 5.
- 749 41. Kubilius, J., Wagemans, J., and Op de Beeck, H.P. (2014). A conceptual framework of
750 computations in mid-level vision. *Front Comput Neurosci* 8, 158.
- 751 42. Friston, K.J., and Kiebel, S. (2009). Predictive coding under the free-energy principle.
752 *Philos T R Soc B* 364, 1211-1221.
- 753 43. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of
754 cognitive science. *Behavioral and Brain Sciences* 36, 181-204.
- 755 44. Gosselin, F., and Schyns, P.G. (2003). Superstitious perceptions reveal properties of
756 internal representations. *Psychol Sci* 14, 505-509.
- 757 45. Smith, M.L., Gosselin, F., and Schyns, P.G. (2012). Measuring Internal
758 Representations from Behavioral and Brain Data. *Current Biology* 22, 191-196.
- 759 46. Nestor, A., Plaut, D.C., and Behrmann, M. (2011). Unraveling the distributed neural
760 code of facial identity through spatiotemporal pattern analysis. *P Natl Acad Sci USA*
761 108, 9998-10003.
- 762 47. Gobbini, M.I., Gors, J.D., Halchenko, Y.O., Rogers, C., Guntupalli, J.S., Hughes, H., and
763 Cipolli, C. (2013). Prioritized Detection of Personally Familiar Faces. *PLoS One* 8,
764 e66620.
- 765 48. van Belle, G., Ramon, M., Lefevre, P., and Rossion, B. (2010). Fixation patterns during
766 recognition of personally familiar and unfamiliar faces. *Front Psychol* 1, 20.
- 767 49. Ramon, M., Vizioli, L., Liu-Shuang, J., and Rossion, B. (2015). Neural microgenesis of
768 personally familiar face recognition. *Proc Natl Acad Sci U S A* 112, E4835-4844.
- 769 50. Zhan, J., Garrod, O.G., Van Rijsbergen, N., and Schyns, P. Modelling Face Memory
770 Reveals Task-Generalizable Representations. *Mendeley Data*
771 <http://dx.doi.org/10.17632/nyt677xwfm.1> (2019)

772

773

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

779 **Competing interests.** The authors declare no competing interests.

780 **Author Contributions.** J.Z., N.VR and P.G.S. designed the research; O.G.B.G. and
781 P.G.S. developed the Generative Model of 3D Faces; J.Z. performed the research;
782 J.Z. and N.VR. analysed the data; and J.Z., N.VR. and P.G.S. wrote the paper.

783 **Figure 1. Reverse correlating mental representations of familiar faces.** (A) Task.
784 Illustrative experimental trial with 6 randomly generated face identities. We instructed
785 participants to use their memory to select the face most similar to a familiar identity (here,
786 'Mary') and then to rate the similarity of the selected face (purple frame) to their memory of
787 'Mary' (purple pointer). (B) Generative Model of 3D face identity (GMF). In its forward
788 computation flow (see identity modelling solid arrow), the General Linear Model (GLM)
789 decomposes a 3D, textured face (e.g. 'Jane' or 'Tom') into a non-identity face shape average
790 capturing the categorical factors of face sex, ethnicity, age and their interactions plus a
791 separate component that defines the identity of the face (illustrated by the 3D shape
792 decomposition; 2D texture, not illustrated, is independently and similarly decomposed). Heat
793 maps indicate the 3D shape deviations that define 'Jane' and 'Tom' in the GMF in relation to
794 their categorical averages. In the reverse flow (see dashed arrow of identity generation), we
795 can randomize the 3D shape identity component (and 2D texture component, not illustrated
796 here), add the categorical average of 'Jane' (or 'Tom') and generate random faces, each with
797 a unique identity that share all other categorical face information with 'Jane' and 'Tom.'

798

799 **Figure 2. Contents of mental representations of familiar faces.** (A) Mental representation
800 of 'Mary' (a typical participant). *Ground truth*: 3D vertex positions deviate both Inward (-) and
801 Outward (+) from the categorical average to objectively define the shape of each familiar face
802 identity. Greyscale values reported on the flanking faces color-code the normalized
803 magnitudes of inward and outward deviations from the categorical average. *Mental*
804 *representation*: Inward and Outward colored faces highlight the individual 3D vertices whose
805 position faithfully deviate from the categorical average in the GMF ($p < 0.05$, two-sided). Blue
806 to red colors represent the normalized magnitudes of their deviations. *2D scatter plots*:
807 Scatter plots indicate the relationship between each vertex deviation in the ground truth
808 (normalized scale on the X-axis) and the corresponding vertex in the memory representation
809 (normalized scale on the Y-axis). The white diagonal line provides the reference of veridical
810 mental representation in the GMF—i.e. a hypothetical numerical correspondence between
811 each shape vertex position in the ground truth face and in the mental representation of the
812 same face. White dots indicate vertices that were not faithfully represented. (B) Mental
813 Representations (group results). Same caption as Figure 2A, except that the colormap now
814 reflects the number of participants ($N = 14$) who faithfully represented this particular shape
815 vertex.

816

817 **Figure 3. NNMF multivariate and compact representations.** A. NNMF representations of
818 faithful 3D vertices across the mental representations of participants. The x-axis heatmap
819 presents each NNMF component, where colors indicate the relative weight of each shape
820 vertex in the component (normalized by maximum weight across components). Boxplots on
821 the y-axis show the loading of each NNMF component on the faithful representations ($N = 14$,
822 one per participant) of each familiar identity ($N = 4$ familiar identities), with colored boxes
823 indicating above 0.1 threshold loading for NNMF components. In boxplots, the bottom (vs. top)
824 edges indicate the 25th (vs. 75th) percentile of the distribution; the whiskers cover the +2.7

825 standard deviation; the larger central circle indicates the median; the outliers are plotted in
826 smaller circle outside the whiskers. **B. Diagnostic and nondiagnostic components for each**
827 familiar identity. Heat maps in the left column show the diagnostic component for each
828 familiar identity; heat maps in the right column show the complementary nondiagnostic
829 components.

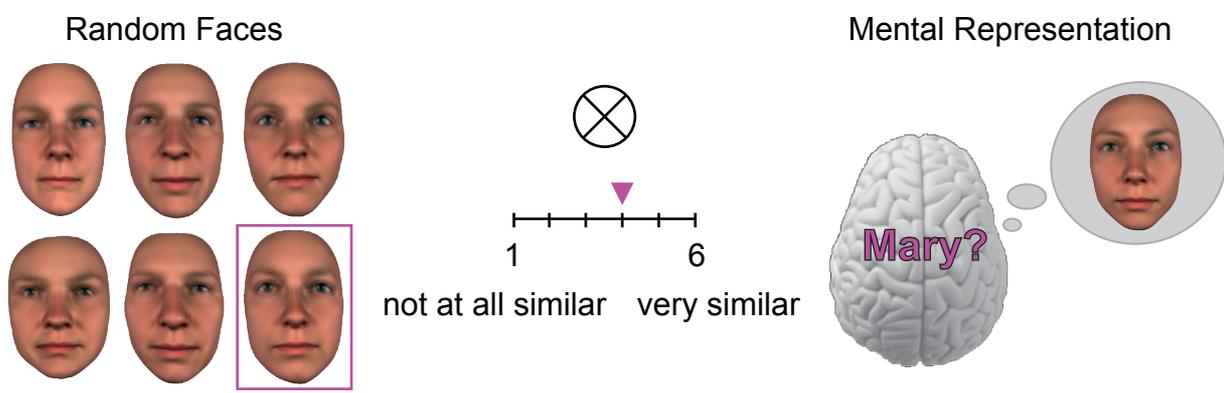
830

831 **Figure 4. Generalization of performance across tasks.** (A) Diagnostic and nondiagnostic
832 Faces. *Left panel:* The red background map shows the multivariate diagnostic components of
833 faithful 3D shape representation of 'Mary'; the grey background map shows the nondiagnostic
834 complement (1 - diagnostic components). *Middle panel:* Faces synthesized with increasing
835 amplification (0.33 to 1.67) of the diagnostic (top) vs. nondiagnostic (bottom) components.
836 *Right panel:* For each synthesized face, we changed its viewpoint (30° left and 30° right), age
837 (80 years old) and sex, shown here for faces synthesized at amplification = 1. (B) Task
838 Performance. For each condition of generalization (row) and familiar identity (column), 2D
839 plots show the median identification performance computed across 12 validators (y-axes) for
840 faces synthesized with the diagnostic (red curves) and nondiagnostic (grey curves) faces, at
841 different levels of amplification of the multivariate components (x-axes). Shadowed regions
842 indicate median absolute deviations (MAD) of identification performance. Abbreviations: Diag
843 = Diagnostic, Nondiag = Nondiagnostic.

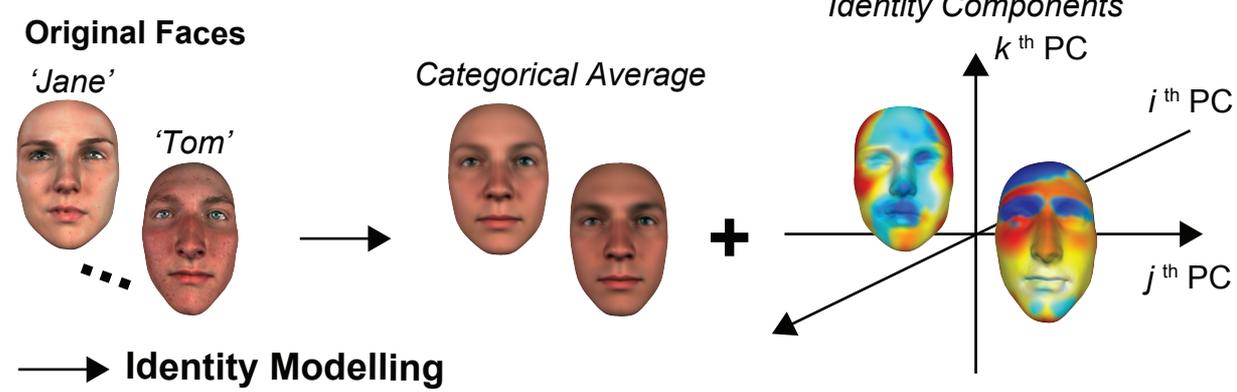
844

845

A. Reverse Correlation Task

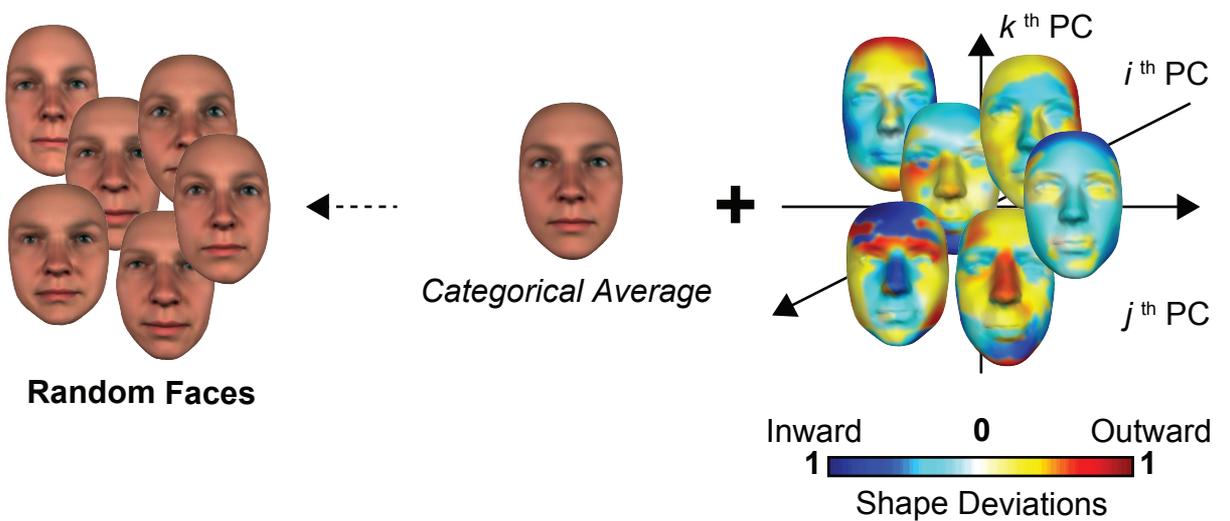


B. Generative Model of 3D Face Identity

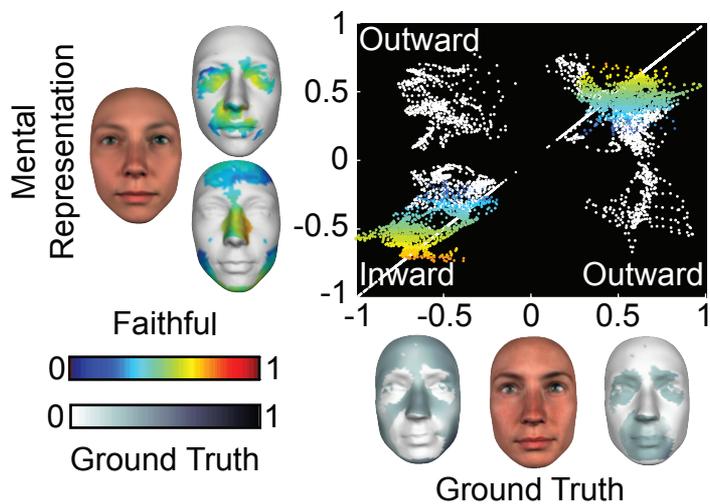


$$\text{Face} = \text{GLM} \left(\begin{matrix} \text{age} \\ \text{sex} \\ \text{ethnicity} \\ \text{interactions} \end{matrix} \right) + \begin{matrix} w_1 & \dots & w_{355} \end{matrix} * \text{PCs}$$

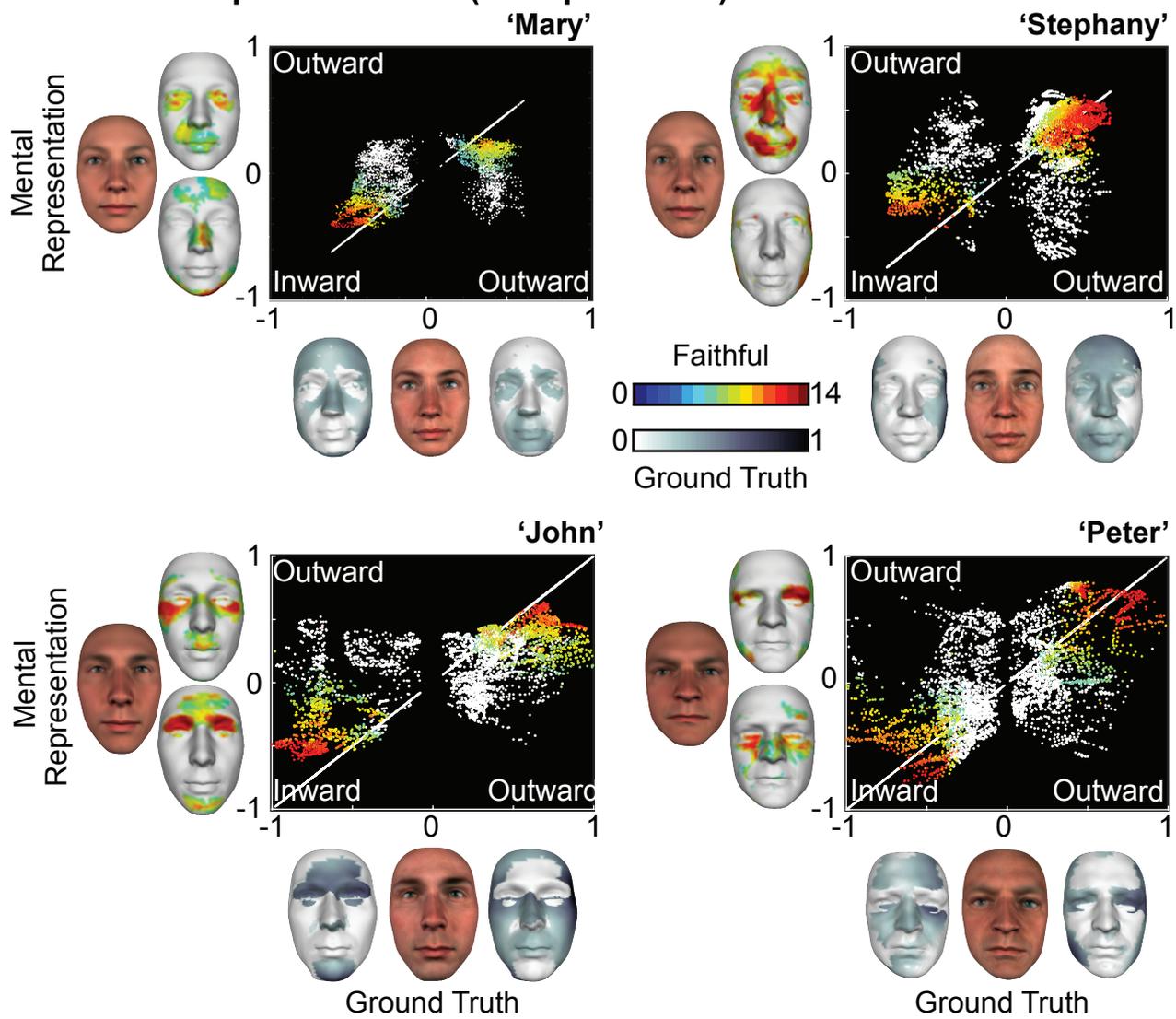
←----- Identity Generation



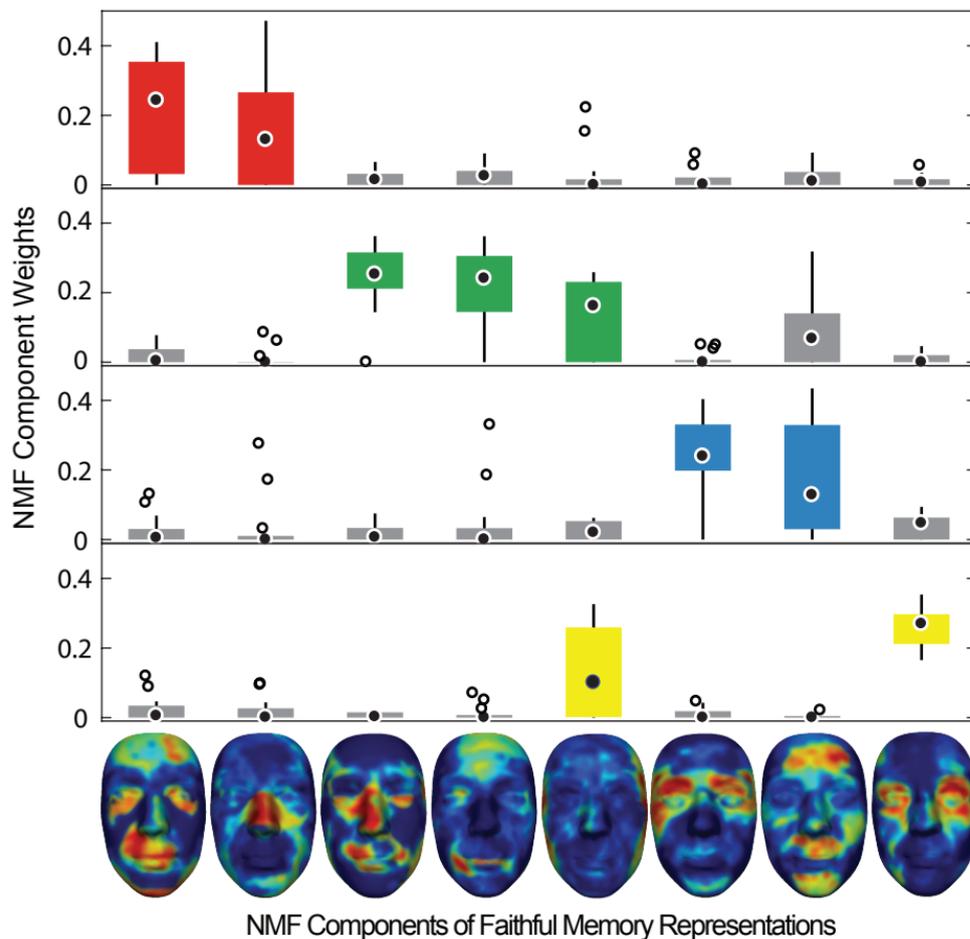
A. Mental Representation of 'Mary' (One Participant)



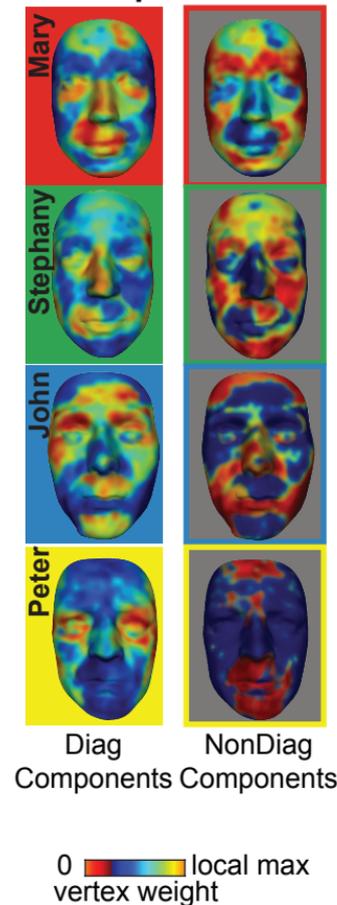
B. Mental Representations (Group Results)



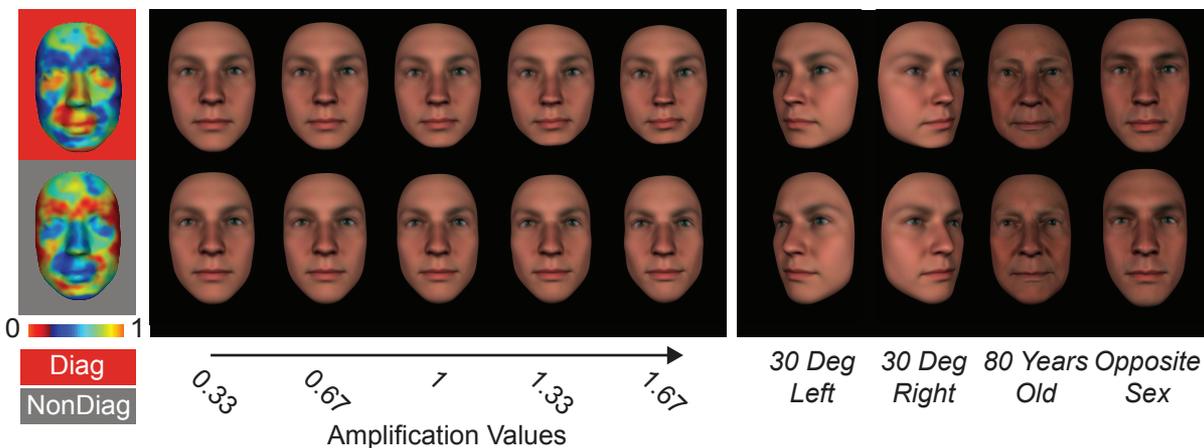
A. Multivariate NMF Representations of Faithful 3D Vertices



B. Diag vs. NonDiag Components



A. Diagnostic and Nondiagnostic Faces



B. Identification Performance of Diagnostic and Nondiagnostic Faces

