UNIVERSITY OF CALIFORNIA SAN DIEGO

Leveraging Computer Vision Face Representation to Understand Human Face Representation

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Chaitanya Krishna Ryali

Committee in charge:

Professor Angela J. Yu, Chair
Professor Garrison W. Cottrell
Professor Sanjoy Dasgupta
Professor Lawrence K. Saul
Professor Edward Vul

2021

The Dissertation of Chaitanya Krishna Ryali is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

This dissertation would not have been possible without the love and support of my family and is dedicated to them.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my advisor, Professor Angela J. Yu, for her guidance, patience and constant support. I learned a lot from Angela and I am very glad to have her in my corner. I also thank my committee members, Prof. Garrison W. Cottrell, Prof. Sanjoy Dasgupta, Prof. Lawrence K. Saul and Prof. Edward Vul for their support and feedback. I am also grateful to all Yu Lab members for many discussions and support.

This dissertation would not have been possible without the support of my friends and family. I thank my friends Dheeraj, Somok, Sankeerth, Dalin, Mahta, Gautam, Tejaswy, Pooja and Shouvik for their support and will cherish the good times we spent together. I thank my parents and sister for their endless patience and love. I also thank my aunt and uncle who have been a second set of parents to me. I thank my wife Pritha for her patience, support and without whom, the world would be a much lonelier place for me.

Chapter 1, in full, is a reprint of the material as it appears in *a*) Ryali CK, Goffin S, Winkielman P, Yu AJ (2020). From Likely to Likable: The Role of Statistical Typicality in Human Social Assessment of Faces. Proceedings of the National Academy of Sciences (PNAS) and *b*) Ryali CK, Yu AJ (2018). Beauty-in-Averageness and its Contextual Modulations: A Bayesian Statistical Account. Advances in Neural Information Processing Systems (NeurIPS). The dissertation author was the primary investigator and author of these papers.

Chapter 2, in full, is a reprint of the material as it appears in Ryali CK, Wang X, Yu AJ (2020). Leveraging Computer Vision Face Representation to Understand Human Face Representation. Proceedings of the Cognitive Science Society Conference (CogSci). The dissertation author was the primary investigator and author of this paper.

VITA

2012 *B. Tech* in Electrical Engineering, Indian Institute of Technology Madras

2015 *M.S.* in Electrical Engineering, University of California San Diego

2021 *Ph. D.* in Computer Science, University of California San Diego

PUBLICATIONS

Ryali CK, Schwab DJ, Morcos AS (2021). Characterizing and Improving the Robustness of Self-Supervised Learning through Background Augmentations. *Under Review*.

Liang Y, **Ryali CK**, Hoover B, Grinberg L, Navlakha S, Zaki M, Krotov D (2021). Can a Fruit Fly Learn Word Embeddings? *International Conference on Learning Representations* (ICLR).

Ryali CK, Hopfield J, Grinberg L, Krotov D (2020). Bio-Inspired Hashing for Unsupervised Similarity Search. *International Conference on Machine Learning* (ICML).

Ryali CK, Goffin S, Winkielman P, Yu AJ (2020). From Likely to Likable: The Role of Statistical Typicality in Human Social Assessment of Faces. *Proceedings of the National Academy of Sciences* (PNAS).

Ryali CK, Wang X, Yu AJ (2020). Leveraging Computer Vision Face Representation to Understand Human Face Representation. *Proceedings of the Cognitive Science Society Conference* (CogSci).

Huang SJ, **Ryali CK**, Liu J, Guo D, Guan J, Li Y, Yu A J (2019). A Model-Based Investigation of the Biological Origin of Human Social Perception of Faces. *Proceedings of the Cognitive Science Society Conference* (CogSci).

Ryali CK, Yu AJ (2018). Beauty-in-Averageness and its Contextual Modulations: A Bayesian Statistical Account. *Advances in Neural Information Processing Systems* (NeurIPS).

Ryali CK, Reddy G, Yu AJ (2018). Demystifying Excessively Volatile Human Learning: A Bayesian Persistent Prior and a Neural Approximation. *Advances in Neural Information Processing Systems* (NeurIPS).

Guan J*, **Ryali CK***, Yu AJ (2018). Computational Modeling of Social Face Perception in Humans: Leveraging the Active Appearance Model. *Technical Report, bioRxiv*.

ABSTRACT OF THE DISSERTATION

Leveraging Computer Vision Face Representation to Understand Human Face Representation

by

Chaitanya Krishna Ryali

Doctor of Philosophy in Computer Science

University of California San Diego, 2021

Professor Angela J. Yu, Chair

Face processing plays a central role in human social life. Humans, including very young babies, readily perform sophisticated computational tasks based on a brief glimpse of a face, such as recognizing individuals, identifying emotional states, and assessing social traits (e.g. attractiveness or trustworthiness). While the accuracy of the last phenomenon, known as physiognomy, is debated, it is consistent among individuals with similar demographic and cultural background and has an impact on real-life decisions such as in dating, employment, education, law enforcement, and criminal justice. A

computational understanding of human face perception is of pressing importance not only for psychology but also for engineered systems that need to anticipate or generate faces with the desired percept and for fairness in law enforcement and the criminal justice system. We leverage ideas and tools from computer vision, metric learning, and information theory to derive scientific insights on how humans represent and perceive faces.

Introduction

Face processing plays a central role in human social life. Humans, including very young babies, readily perform sophisticated computational tasks based on a brief glimpse of a face, such as recognizing individuals, identifying emotional states, and assessing social traits (e.g. attractiveness or trustworthiness). While the accuracy of the last phenomenon, known as physiognomy, is debated, it is consistent among individuals with similar demographic and cultural background and has an impact on real-life decisions such as in dating, employment, education, law enforcement, and criminal justice. A computational understanding of human face perception is of pressing importance not only for psychology but also for engineered systems that need to anticipate or generate faces with the desired percept and for fairness in law enforcement and the criminal justice system. We leverage ideas and tools from computer vision, metric learning, and information theory to derive scientific insights on how humans represent and perceive faces. Below, we provide a brief overview of the chapters.

Chapter 1

Humans readily form social impressions, such as attractiveness and trustworthiness, from a stranger's facial features. Understanding the provenance of these impressions has clear scientific importance and societal implications. Motivated by the efficient coding hypothesis of brain representation, as well as Claude Shannon's theoretical result that maximally efficient representational systems assign shorter codes to statistically more

typical data (quantified as log likelihood), we suggest that social "liking" of faces increases with statistical typicality. Combining human behavioral data and computational modeling, we show that perceived attractiveness, trustworthiness, dominance, and valence of a face image linearly increase with its statistical typicality (log likelihood). We also show that statistical typicality can at least partially explain the role of symmetry in attractiveness perception. Additionally, by assuming that the brain focuses on a task-relevant subset of facial features, and assessing log likelihood of a face using those features, our model can explain the "ugliness-in-averageness" effect found in social psychology, whereby otherwise attractive, inter-category faces diminish in attractiveness during a categorization task.

Chapter 2

We leverage various computer vision techniques, combined with human assessments of similarity between pairs of faces, to investigate human face representation. We find that combining a shape- and texture-feature based model (Active Appearance Model) with a particular form of metric learning, not only achieves the best performance in predicting human similarity judgments on held-out data (both compared to other algorithms and to humans), but also performs better or comparable to alternative approaches in modeling human social trait judgment (e.g. trustworthiness, attractiveness) and affective assessment (e.g. happy, angry, sad). This analysis yields several scientific findings: (1) facial similarity judgments rely on a relative small number of facial features (8-12), (2) race- and gender-informative features play a prominent role in similarity perception, (3) similarity-relevant features alone are insufficient to capture human face representation, in particular some affective features missing from similarity judgments are also necessary for constructing the complete psychological face representation.

Chapter 1

From Likely to Likable: the Role of Statistical Typicality in Human Social Assessment of Faces

1.1 Introduction

Humans readily form social impressions, such as attractiveness and trustworthiness, from a brief glance of a stranger's face [2, 133, 26, 119]. Even infants show a looking preference for faces that are judged by adults to be more attractive [108] or more trustworthy [44]. While the accuracy of these social judgments is an area of active research [126, 119], such social impressions clearly exert a powerful influence on real-life decisions [75, 119], whether choosing a life partner, assessing eye witness testimony, interviewing a job candidate, or choosing whom to befriend. One of the most robust and intriguing findings in the study of social judgment of faces is the so-called "beauty in averageness" (BiA) effect, whereby blends of two or more face images are generally perceived to be more attractive than the "parent" face images [25, 54]. A number of qualitative hypotheses have been put forth to explain the provenance of BiA, such as a human preference for symmetry [47, 91] or lack of blemishes [55]. However, symmetry and blemishes appear at most to provide an incomplete explanation of the BiA effect, as controlling for these

factors does not eliminate BiA [48]. Alternatively, it has been suggested that humans have a preference for highly prototypical stimuli over more unusual stimuli [134], possibly as a cue to mate value or reproductive health [115, 116]. However, humans exhibit BiA effects not only for strangers' faces, but also for a variety of natural and artificial object categories such as dogs, birds, butterflies, fish, automobiles, watches, and even synthetic dot patterns [36, 37, 134]. Beyond attractiveness, trustworthiness perception has also been found to be more positive for faces that are more typical [110]. Other recent studies have shown that attractiveness and trustworthiness perception are culturally dependent [109], and can be rapidly modified via experimental exposure of specific types of faces [21]. Altogether, these results suggest that there may be a general human social preference for more typical-looking objects, in particular faces, and this effect depends on rather general cognitive mechanisms beyond those specific to beauty or mating, or even specific to facial features.

In this work, we propose a *statistically grounded* model for human "liking" of more typical-looking faces (and other objects) [34, 110, 41]. We suggest that the brain internally represents the statistical distribution of faces based on experience, either implicitly or explicitly. The perceived appeal of a face is, at least in part, *monotonically driven* by a measure of "typicality" of the face with respect to the face distribution, specifically in terms of the logarithm of the likelihood (LL) of the face under the distribution. To differentiate our statistical definition of typicality from previous non-statistical proposals of the link between typicality and attractiveness [34, 110, 41], we refer to our measure as *statistical* typicality. Intuitively, as long as the face distribution is unimodal (e.g. approximately Gaussian, as depicted in Figure 1.1A), the blend (average) of two "parent" faces will generally be closer to the mean and thus have higher LL than the "parent" faces, resulting in BiA [25, 54, 110].

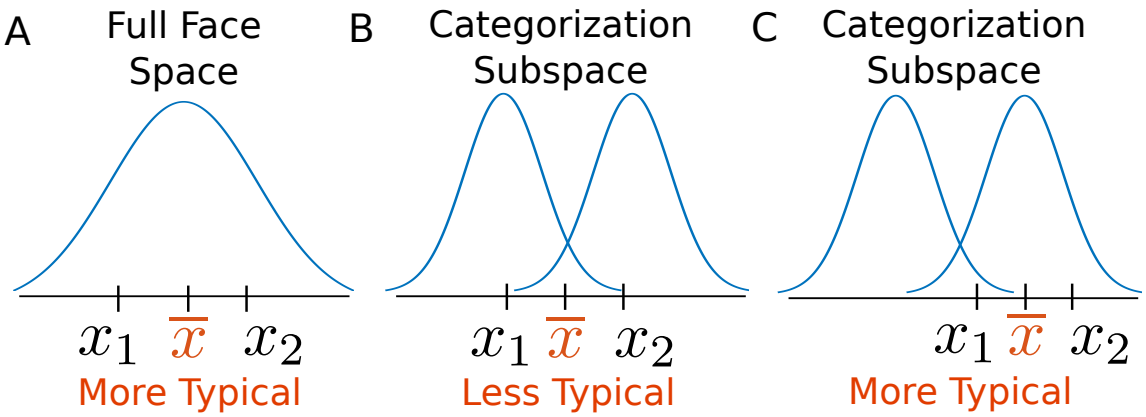

Figure 1.1. Schematic illustration. (A) In a unimodal distribution, the blend (average) of two face stimuli tends to have higher statistical typicality than the "parent" stimuli. In a bimodal distribution, (B) the blend of two "parents" from two different modes tends to have lower statistical typicality than (C) when the "parents" come from the same mode.

Methodologically, we model faces as points in a vector space ("face space") [124] defined by their feature representation in a widely used computer vision model [22, 13, 123], and use a demographically-balanced, publicly available face dataset, Chicago Face Database (CFD) [60], to model the statistical distribution of faces in this feature space. As we will show, attractiveness ratings from CFD correlate with LL of these faces under the estimated density function, and simulations based on CFD faces broadly exhibit empirically observed BiA effects. Moreover, in a novel experiment, we find that the model-estimated LL of face stimuli linearly predicts not only perceived attractiveness but also trustworthiness, dominance, and valence – all traits thought by social psychologists to be particularly fundamental in face-based social judgments [119]. Additionally, we demonstrate that even the well-known role of symmetry in attractiveness perception [47, 91] may be a special case of statistical typicality, as symmetry correlates with LL among CFD faces, and experimental techniques that increase face symmetrization also increase LL.

If human liking of a face is indeed related to statistical typicality, then it makes

a particularly interesting prediction when the data come from a multi-modal mixture distribution: if two faces are drawn from two different mixture components, then their blend may well have lower LL, and thus attractiveness, than the "parent" faces (Figure 1.1B). In fact, this dovetails with empirical findings that bi-racial and bi-gender blends are perceived to be significantly less attractive than same-race or same-gender blends, but only in the race- [38] or gender-categorization condition [78], respectively (outside these specific conditions, bi-racial and bi-gender blends still exhibit BiA). Like BiA, this "ugliness-in-averageness" (UiA) effect has also been observed with non-face stimuli [127]. Within our framework, we explain UiA as follows. During the categorization task, attention focuses on a task-relevant subspace of facial features (e.g. the features that best discriminate gender in the gender categorization task) [17, 138, 12, 31, 101, 16]. Furthermore, attractiveness perception is linked to LL defined within this task-relevant subspace. Because a cross-category blend tends to live in between categories, it has relatively low LL and low attractiveness (Figure 1.1B). In contrast, a same-category blend will live in the same mode as the "parent" faces and thus exhibit BiA (Figure 1.1C). We will use simulations of the model to show how experimentally observed UiA effects arise directly from the LL of faces when the distribution is restricted to the categorization-relevant subspace of facial features. As a stronger test of our model, we will show that model-predicted LL of *individual* face images within the categorization-relevant subspace correlates significantly with subject-reported attractiveness rating of those faces [78].

One may well ask why the there should be an affective signal in the brain related to statistical typicality, and why statistical typicality should take the form of log likelihood (LL). For theoretical motivation, we appeal to Claude Shannon's classical result [103] that a maximally efficient coding system (minimizing the average amount of code needed to represent the data) should assign to each data point a code length that is exactly

proportional to its negative LL, i.e. shorter codes for more frequently encountered data. For example, Morse code is a fairly efficient code, consisting of short-duration dots and long-duration dashes, that assigns the shortest code to the most frequently used letter in English (e.g. E is 1 dot) and much longer codes to infrequently used letters (e.g. Q and Y are both 3 dashes and 1 dot, arranged in different orders). Analogously, the classical "efficient coding hypothesis" in neuroscience proposes that information representation and processing in the brain are highly efficient [3, 59, 61], such that less neuronal response is allocated to encode more probable stimuli [4, 14, 3, 56, 131, 82]. Supporting efficient coding of faces in the brain, face-responsive areas (amygdala, fusiform face area, occipital face area, face-selective regions of the posterior superior temporal sulcus) have been found to exhibit the lowest fMRI BOLD response to the most average-looking face, and increasingly higher responses to less typical faces [98, 66].

To attain and maintain coding efficiency in the brain, it is helpful for inefficiently coded stimuli to be aversive, so as to encourage representational updating that will minimize long-term coding cost averaged across observed data [140, 24]. Interestingly, reducing the average negative LL of observed data under the assumed model, known as *cross-entropy minimization*, is also a popular and effective tool for data modeling in modern machine learning and artificial intelligence [8]. In the neuroscience literature, there is empirical evidence that face stimuli whose representation in the brain requires greater neural activity is aversive. For example, an EEG event-related potential (ERP) study found that more attractive and average-looking faces evoke a weaker (posterior N170) response than unattractive (and atypical) faces, resulting from engaging fewer neural resources in the former case [122]. There is also evidence that specifically in the UiA context, there are neural responses whose increase leads to less social "liking" of a face [114, 49]. For example, a late positive potential (LPP) has been found in EEG studies

to specifically increase for emotionally ambiguous faces (e.g. happy vs. angry), but only in an emotion categorization context [114, 49]; moreover, the magnitude of the LPP predicts how strong the UiA effect is (how untrustworthy the ambiguous face is compared to a control task) across individual observers [49]. Intriguingly, this LPP has been found to be localized around the anterior cingulate cortex [114, 49], which we previously found to encode an unsigned prediction error, highly related to statistical atypicality, in a Bayesian predictive coding context [42].

In summary, there is a diverse range of empirical and theoretical results that coherently support efficient coding as a theoretical motivation for the influence of statistical typicality on social liking of faces. Efficient coding also provides a theoretical justification for why typicality should be measured as LL. However, this paper primarily focuses on the relationship between LL and social liking of faces, and thus its main findings stand independently of the extent of the explanatory role played by efficient coding.

In the following, we demonstrate how our model assumptions and predictions are validated by a combination of face data, statistical modeling, human behavioral data, and model simulations. We begin with an analysis of BiA effects, then proceed to UiA.

1.2 Results

As discussed in the introduction, we hypothesize BiA arises because the blend of two faces tends to be closer to the mean of the distribution and have higher LL than the "parent" faces (Figure 1.1A). To examine the statistical distribution of faces (see Methods and SI), we utilize a demographically balanced, publicly available dataset (Chicago Face Database, CFD) of 597 face images [60], and embed them into the latent feature representation of a well-established computer vision model, the Active Appearance Model (AAM) [22, 13, 123]. AAM has previously been used to model human face representation [33, 41],

and AAM feature dimensions (linear axes) appear to be encoded linearly by face-selective neurons in the primate brain [9].

1.2.1 Simulation: Beauty-in-Averageness

To check whether statistical typicality (LL) of CFD faces and their blends can reproduce experimentally observed BiA effects, we need a parametric density model of faces. Due to the relatively small number of CFD faces (597) compared to the large number of AAM features (90), we fit a multivariate normal distribution to the CFD data, although we also find that LL is highly correlated whether we fit a single Gaussian or a mixture of Gaussians corresponding to demographic subgroups (e.g. genders, see SI), and the two models correlate with CFD attractiveness ratings similarly well (see SI). Moreover, taking 500 random 1-d projections of the CFD data in the AAM feature space (see Figure 1.2A for an example projection), we find that normality cannot be rejected (Anderson Darling test, significance level $\alpha = 0.05$) in all but 38 dimensions (7.6%), not much higher than the 5% we would expect if every dimension was truly normal (in contrast, if we simulate random projections of data sampled from a distribution that was bimodal in every dimension, then 60% of them reject normality, see SI). In fact, we do expect a small number of dimensions to be multi-modal due to natural demographic clustering of data (e.g. due to gender or race), an important point we will return to in the UiA analysis. In the following, we use the CFD-estimated multivariate Gaussian density for all the BiA-related analyses.

Our simulations (see SI) indicate that the blend of two randomly sampled CFD faces has higher LL (thus presumably higher attractiveness) than "parent" face images (2-sample t-test, $p < 0.0001$), consistent with experimental findings [54]. Moreover, as the number of "parent" faces increases in a blend, the blended face increases in LL (and thus attractiveness) monotonically (ANOVA test, $p < 0.0001$; t-test comparing 2-face blend to

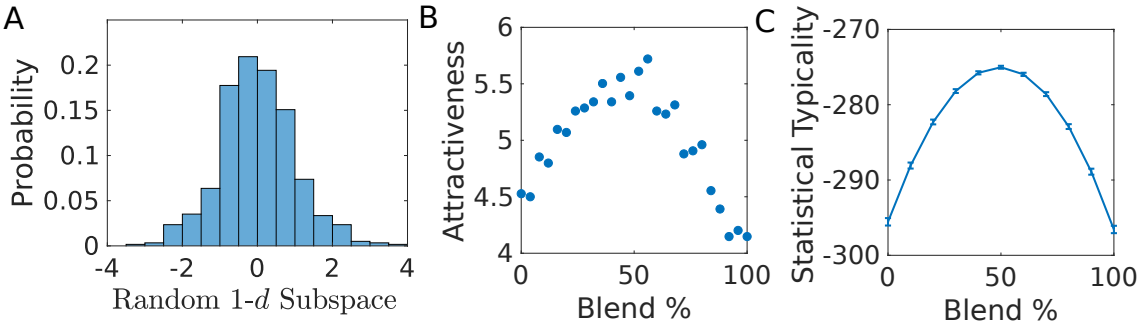

Figure 1.2. A. Empirical face distribution projected in a random direction is unimodal and bell shaped. B. BiA: humans perceive more "evenly" blended faces as more attractive (data from [38]). C. Simulated statistical typicality captures a similar trend as seen in data (B). 500 pair of faces were randomly selected from CFD, each pair's coordinates were averaged in varying proportions, and LL of each blend was recorded and averaged across pairs. Error bars: s.e.m.

32-face blend, $p < 0.0001$), also consistent with experimental findings [54]. Relatedly, when two faces are blended in different proportions, from 100% of one face to 100% of the other face, LL has a characteristic inverted U-shape, similar to experimental findings [38] (Figure 1.2B;C).

1.2.2 Experiment: Social Liking Depends Linearly on Statistical Typicality

If statistical typicality modulates the affective experience of perceptual stimuli, then it should not only be restricted to the perception of attractiveness, but also influence other desirable attributes such as trustworthiness [110]. Here, we report results from an experiment (see Methods), in which we asked subjects to rate face images (see Figure S2 for example stimuli) for attractiveness, trustworthiness, dominance, and valence (how "positive" a face appears, see Methods) – traits thought by social psychologists to be the most fundamental in face-based social trait perception [119]. We find that face image ratings (averaged across subjects) of all four traits increase monotonically with LL of the face stimuli under the estimated multivariate normal density model (attractiveness:

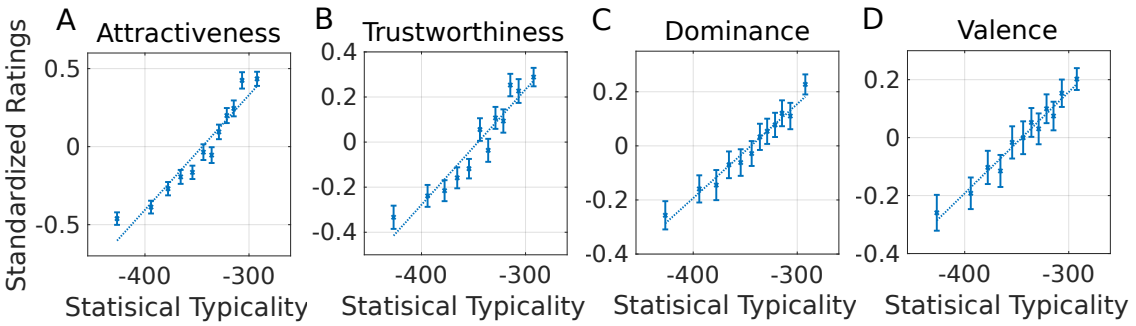

Figure 1.3. Trait rating increases linearly and monotonically against statistical typicality (LL) for all four traits. Data binning ensures equal number of samples in each bin. Linear regression line (using binned data) is superimposed for visualization. See text for correlation coefficients for raw data. Error bars: s.e.m. over samples in each bin.

Pearson $r = 0.386, p < 0.0001$, trustworthiness: $r = 0.268, p < 0.0001$, dominance: $r = 0.196, p < 0.0001$, valence: $r = 0.171, p < 0.0001$). The correlation coefficients indicate that the effect is strongest for attractiveness, followed by trustworthiness, dominance, and finally valence. Figure 1.3 shows that this relationship is highly *linear* for all four traits (binned data: Pearson $r = 0.96, 0.96, 0.98$, and 0.98, respectively). At the individual level, we also find that statistical typicality correlates with ratings for a significant fraction of participants for each trait (binomial test, $p < 0.0001$ for each trait). Analogous to group-level analysis, more individuals show significant correlation (at $\alpha = 0.05$ significance level) for attractiveness (75% of participants) and trustworthiness (55%) than for dominance (40%) and valence (32.5%) (see Figure S4 for histogram of individually significant c.c. for each trait).

1.2.3 Experiment: Linear vs. Quadratic Models of Attractiveness

One corollary of the statistical typicality account of attractiveness is that, if the distribution of faces is Gaussian over the underlying face feature space, then LL is a particular parameter-free quadratic function of the underlying face features (see *SI* for derivation). This would be consonant with previous work showing that a freely fitted

quadratic model of human perception of attractiveness improves over a pure linear model [99, 120]. However, previous work did not set forth a particular hypothesis for the quadratic component, or analyze it in relation to statistical properties of the face distribution.

Here, we make a more refined claim that the quadratic component is precisely LL. Because LL of a multivariate normal distribution falls off monotonically from the mean face along any particular dimension, it can be thought of as a measure of "averageness" [51, 77, 99, 21]. In contrast to our finding, previous studies found a very small effect of "averageness" on facial attractiveness, which may have resulted from having utilized different distance measures, such as Euclidean distance [41] or standardized Euclidean distance (normalized by std in each dimension) [99] – in contrast, LL can be viewed as Mahalanobis distance (differing from Euclidean distance by taking into account the correlation between shape and texture features, see SI). A larger quadratic effect might also be present in our study because the face stimuli were designed to vary systematically in LL (see Methods and SI). More importantly, face stimuli in our experiment vary among all (90) face feature dimensions, where Euclidean distance, standardized Euclidean distance, and LL all differ from each other non-trivially, while previous studies [21] used stimuli that live on a single feature dimension, along which the three distance metrics are all confounded.

We hardly expect statistical typicality to be the sole contributor to attractiveness. Previous work provided evidence that human perception of attractiveness, as well as other traits such as trustworthiness and dominance, has both linear and squared dependence on the face feature space [99, 120]. Here, we adapt this idea and jointly fit a multiple linear regression (MLR) model consisting of all the linear terms $(x_1, \ldots, x_k)$, along with a LL term pre-multiplied by a coefficient – we dub this the "linear+LL" model (see

Figure 1.4A for an illustration). For comparison, we also fit a regression model with only linear predictors ("linear model"), and another with all the linear and squared terms ("quadratic" model [99, 118]). Because the three models have vastly different numbers of free parameters (linear: 90 features, 1 offset; linear+LL: 90 features, 1 LL term, 1 offset; quadratic: 90 features, 90 squared features, 1 offset), it is inappropriate to compare them directly on the training data. Instead, we do 10-fold cross-validation: we train the model on 90% of the dataset, then compute the correlation coefficient (c.c.) between the predicted trait rating of held out (10%) test data and the human average ratings on those faces; we then rotate the partitioning of the data into training data and test data and repeat the process, until every face has served once as a test data point. Across the ten folds, Figure 1.4B shows that linear+LL consistently outperforms the quadratic model for each of attractiveness, trustworthiness, dominance, and valence (paired one-sided t-test: $p = 0.0412, p = 0.0003, p = 0.089, p_{=}0.0003$), as well as the linear model (paired one-sided t-test: $p < 0.0001, p < 0.0001, p < 0.0001, p = 0.0003$). Alternatively, we find that linear+LL has the best (lowest) Bayesian information criterion (BIC) scores compared to the linear and the quadratic models (see Figure S3). These results indicate that by using a particular parameter-free form of a quadratic function, provided directly by the probability density of the data, we can account for social perception across a number of traits better than a more complex model that freely fits all possible squared terms. For brevity, we refer to the vector of regression coefficients for the linear terms in linear+LL as the Linear Trait Axis (LTA) for that trait in the remainder of the paper.

A Common Quadratic Component and Distinct Linear Components

The astute reader might have noticed that the quadratic component (LL) is shared by all social traits, while the linear component (LTA) is fit to each trait individually. To investigate this further, we perform the following analysis: using 10-fold cross-validation,

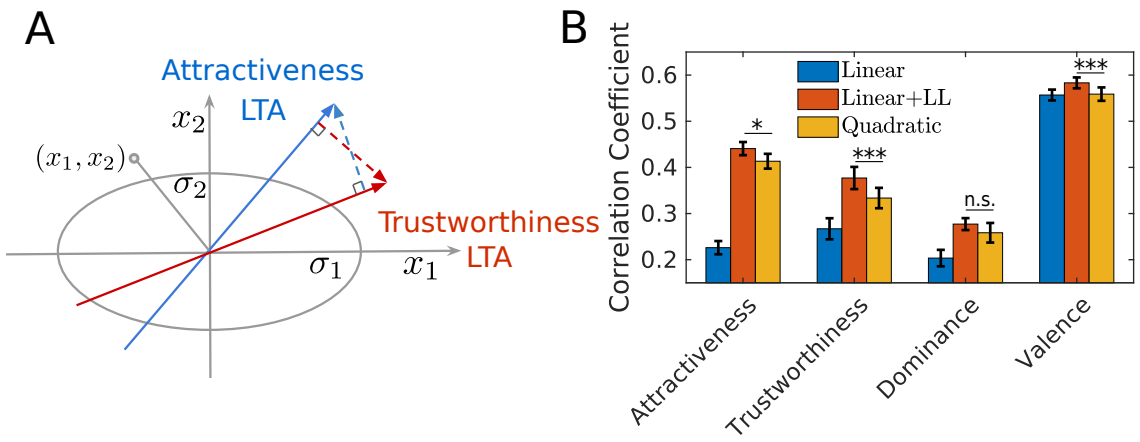

Figure 1.4. Model illustration and comparison. A. We use a 2-*d* simplified illustration to visualize the linear+LL model (in actuality there are 90 dimensions). Oval: equi-LL contour of an axis-aligned normal density function, with σ_1 and σ_2 being the standard deviations (std) along the two feature dimensions x_1 and x_2 (estimated normal density may not be axis-aligned). Linear+LL model of attractiveness is a linear combination of the quadratic component related to negative LL and a linear component indicated by the "attractiveness LTA" (blue); similarly, linear+LL models trustworthiness as a linear combination of the negative LL quadratic component and a linear "trustworthiness LTA" component (red). -LL of a face situated at (x_1, x_2) is proportional to the sum of the square of its distance along each coordinate divided by the std of that coordinate – in particular, x_2 is given a greater weight than x_1 in this example, because the data is more tightly distributed for x_2 than for x_1. The component of "attractiveness LTA" orthogonal to "trustworthiness LTA" (blue dashed line) is an axis along which attractiveness ratings should look linear, while trustworthiness ratings should look quadratic, as was found in [110]. However, the component of "trustworthiness LTA" orthogonal to "attractiveness LTA" (red dashed line) is an axis along which trustworthiness ratings should look linear, while attractiveness ratings should look quadratic. B. Correlation coefficient between predicted and actual ratings on held-out data, averaged over 10 folds. Linear+LL consistently outperforms the quadratic and linear models. Error bars are s.e.m over folds.

we find that the LTA for each trait makes better prediction about the same trait on held-out test data than the LTA for any other trait (paired t-test, $p < 0.0001$, see Table 1.1), except for the trait "trustworthiness", which is slightly better predicted by the LTA for valence ($r = 0.307$) than for trustworthiness ($r = 0.270$), but this is not statistically significant ($p = 0.384$). In particular, it is apparent that dominance actually has quite different linear explanatory factors than the other traits, but nevertheless it shares the same quadratic, typicality-driven factor as the rest. These results also shed new light on a recent finding that typicality affects trustworthiness perception, and the suggestion that BiA may be an indirect phenomenon due to the high correlation between trustworthiness and attractiveness [110]. What we find is that indeed trustworthiness and attractiveness are highly related: not only do they have similar linear components (their LTAs predict each other quite well, though not as well as each trait's own LTA, see Table 1), but they also share a common typicality-driven quadratic component (Figure 1.3). In our hands, trustworthiness and attractiveness are individually driven by statistical typicality: in fact, statistical typicality (LL) correlates more strongly with attractiveness ($r = 0.39$) than trustworthiness ($r = 0.27$). To further confirm this, we regress out trustworthiness ratings of faces and find that residual attractiveness is still strongly correlated with statistical typicality ($r = 0.33, p < 0.0001$); likewise, residual trustworthiness, once having regressed out attractiveness, still strongly correlates with statistical typicality ($r = 0.18, p < 0.0001$). In other words, attractiveness and trustworthiness do not "inherit" their dependence on typicality from each other, but are instead each driven directly by statistical typicality. Given that both traits have both a (shared) quadratic component and a (unique) linear component component (Figure 1.4), not only should one be able to find an axis along which trustworthiness looks quadratic while attractiveness looks linear (as found in [110]), but it should also be possible to find an axis along which attractiveness looks

Table 1.1. Trait-Specific LTAs. Each trait is generally better predicted by its own LTA than other LTAs, measured by the correlation coefficient between model predictions on ratings of held-out faces and predictions by all LTAs, averaged across the 10 folds of the cross validation.

	Trait Predicted			
LTA	Attractiveness	Trustworthiness	Dominance	Valence
Attractiveness	<u>0.23</u>	0.23	−0.10	0.35
Trustworthiness	0.21	0.27	−0.12	0.49
Dominance	−0.10	−0.14	<u>0.20</u>	−0.24
Valence	0.20	<u>0.30</u>	−0.13	<u>0.56</u>

quadratic while trustworthiness looks linear (see Figure 1.4A).

1.2.4 Simulation: Symmetry and Statistical Typicality

Revisiting the role of symmetry in attractiveness perception [47, 91], we suggest the possibility that symmetry elevates attractiveness perception at least in part through statistical typicality. Figure 1.5A shows that the statistical typicality (LL) of CFD face images are negatively correlated (Pearson $r = -0.373$, $p < 0.0001$) with shape asymmetry (Euclidean distance between landmarks and their mirror image). As a whole, a large majority of CFD faces (98.3%, binomial test, $p < 0.0001$) increase LL after shape symmetrization (Figure 1.5B, symmetrization technique similar to [46], see SI), and more *atypical* faces benefit more in LL from shape symmetrization (regression coefficient of LL(symmetrized face) - LL(original face) against -LL(original face) is significantly greater than 0, $p < 0.0001$; regression coefficient $b = 1.21$, 95% CI $[1.12, 1.30]$). As a whole, shape symmetrization increases LL of CFD faces by 1.01 in units of standard deviation of the original empirical LL distribution of CFD faces (paired t-test, $p < 0.0001$; see Figure 1.5C). If texture features are left-right symmetrized in addition to shape features (not typically done in experiments), our model predicts an even bigger effect of symmetrization

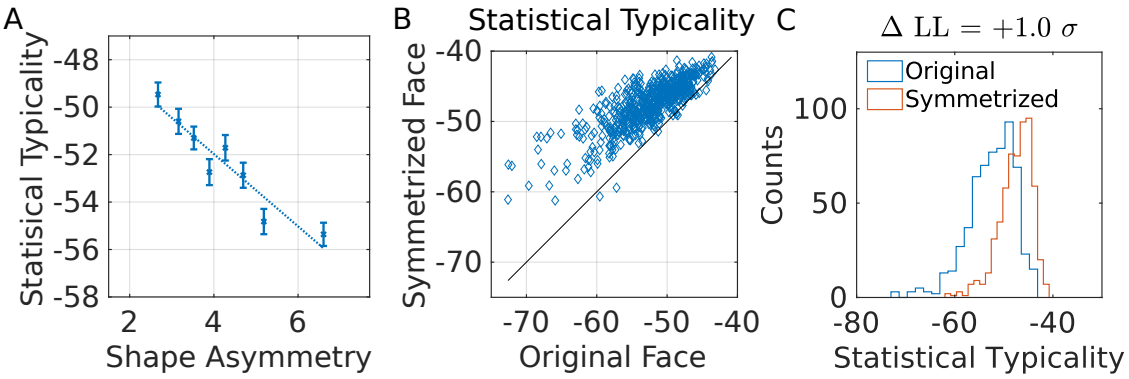

Figure 1.5. Symmetry and statistical typicality. (A) LL negatively correlates with shape asymmetry. (B,C) Shape symmetrization increases LL of most face images. (A) Data binning ensures equal number of samples in each bin. Linear regression line (using binned data) is superimposed for visualization. Error bars: s.e.m. over faces in each bin.

on LL (see Figure S8). Moreover, similar to human data [48], our simulations show that blends of (shape) symmetrized faces still exhibit BiA (larger LL than their "parent faces"), though the difference is slightly smaller than for blends of original (non-symmetrized) faces (see Figure S5), indicating that symmetry has a significant but partial contribution to both experimentally observed attractiveness and model-predicted statistical typicality.

1.2.5 Ugliness-in-Averageness (UiA)

As discussed in the introduction, we hypothesize UiA arises because attentional mechanisms focus on the task-relevant feature subspace, such that statistical typicality and subjective liking are both assessed within this subspace instead of the full (original) face space. If the data distribution projected into this subspace is bimodal, then statistical typicality and subjective liking of an average between two samples from two different modes should be lower than if the "parent" faces were within the same mode (Figure 1.1B).

When we project CFD faces into the 1-d subspace that best discriminates race (Caucasian versus Asian [38, 78], found by regularized linear discriminant analysis, see SI), we find the face distribution to be indeed bimodal (Figure 1.6C). Likewise, projecting CFD

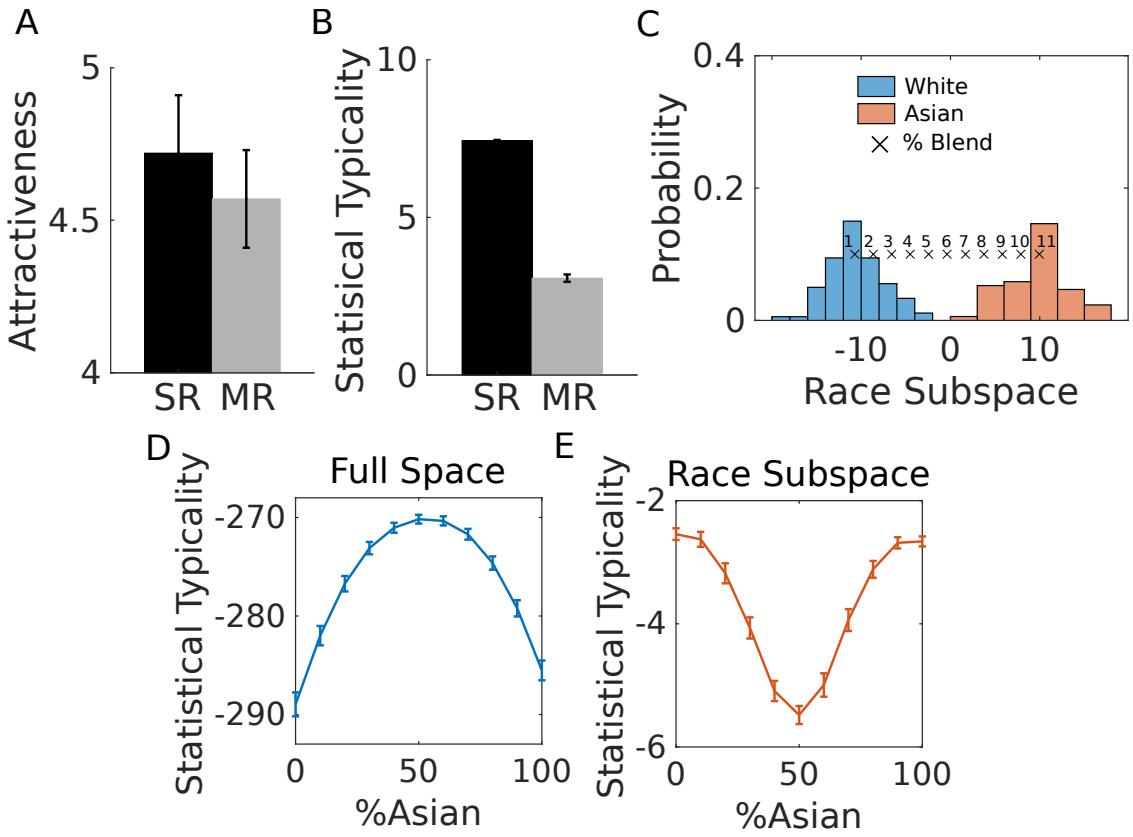

Figure 1.6. Simulation: UiA due to bimodality in the race-informative subspace. A. *Data*, adapted from [38]: single-race (Asian-Asian, White-White, SR) face blends are rated as more attractive than mixed-race (Asian-White, MR) face blends, when a race categorization task precedes attractiveness rating (two sample t-test, $p < 0.05$). Error bars: 95% CI. B. *Model*: statistical typicality of mixed-race blends is lower than single-race blends in the 1-d race-informative subspace. C. Empirical distribution of white and Asian faces [60] projected into the race-informative subspace is bimodal. X: mean location of face images (60 total) for each % of blend, i.e. 1: 100% white and 0% Asian, ..., 11: 0% white and 100% Asian. D. Model-predicted statistical typicality (LL) for actual face images (binned by racial blend %) in the full face space. E. Model-predicted LL for face images (same as in D) in the race-informative subspace. Error bars in B, D, E: s.e.m. over simulated face stimuli.

faces into the gender-informative 1-*d* subspace also exhibits clear bimodality (Figure 1.7B, see SI). It may seem odd that the face distribution is both approximately Gaussian and a mixture distribution. The reason is that the mixture components only appear as distinct components (multi-modal) when viewed from a small number of very particular feature dimensions (e.g. important for discriminating race or gender), but still Gaussian in the great majority of dimensions (the components highly overlap).

Simulation: UiA Reproduced by LL in Task-Relevant Subspace

Human subjects have been found to rate single-race face blends more attractive than mixed-race face blends when they are required to categorize race before rating attractiveness (Figure 1.6A, [38]). We simulate the effect of this race-categorization task on statistical typicality as follows. First, we randomly sample face images from CFD [60] to create 100 single-race (Asian-Asian, White-White) and 100 mixed-race (Asian-White) blends. Then, we project all face blends into the race-informative 1-*d* subspace. The model-predicted statistical typicality for each face blend is its LL under the *a posteriori* most probable gender category (see SI), consistent with a body of work showing that people often use category membership to predict features of and reason about members of a category [92, 64, 63, 11]. We find that model-predicted statistical typicality reproduces the empirically observed UiA effects (Figure 1.6B).

We can make more refined predictions by smoothly varying the % of blending in each pair of faces. We first randomly draw 60 Asian and white face images (with replacement) from the face dataset [60], and then blended them at 10% increments, from 100% of the white face to 100% of the Asian face, thus producing 11 morphs of each pair (see Figure 1.6C). We find that average statistical typicality, as a function of % blend, has an inverted U shape (BiA) relative to % racial blend in the original space (Figures 1.6D) or a random subspace (see Figure S11), but a U shape (UiA) in the race-informative subspace

(Figures 1.6E). Analogous simulations corresponding to the case of gender categorization [78] exhibit a similar pattern (see Figure S9).

Image-Level Comparison: Model vs. Data

As a more stringent test of our model, we investigate the relationship between individual face's attractiveness ratings and model-predicted statistical typicality in the attended, task-relevant (LDA) subspace. We re-analyze the stimuli and behavioral data from the gender categorization study [78], in which subjects rated the attractiveness of blends from male and female "parent" faces, in different proportions (10% increment), under either the control condition (no gender categorization), or the experimental condition (following gender categorization). It was found that attractiveness ratings in the experimental condition, after subtracting out their ratings in the control condition, exhibits UiA: more even gender blends are relatively more negatively affected by the gender categorization (Figure 1.7C [78]). When we embed the blended face stimuli into our AAM face space, and assess statistical typicality in the gender-informative (via LDA) subspace, we also observe a similar UiA effect (Figure 1.7D), in contrast to the BiA effect when statistical typicality is assessed in the full face space (Figure 1.7E). In fact, at the individual face level, we find that the statistical typicality of blended face images significantly correlates with (subtractively normalized) attractiveness rating in the experimental condition ($r = 0.28$, $p = 0.0026$, see SI Figure S6 for a scatter plot). This relatively high correlation coefficient at the individual face level is impressive, because there are clearly other determinants of facial attractiveness besides typicality, such as perceptual and conceptual priming, contrast, clarity, sexual dimorphism, etc. [134, 99, 46]. Importantly, the entirety of this predictive power comes from theoretical considerations of the relationship between LL (based on a density model fit to a different dataset, CFD [60]) and liking - no free parameter was estimated using the ratings or stimuli from the experiment.

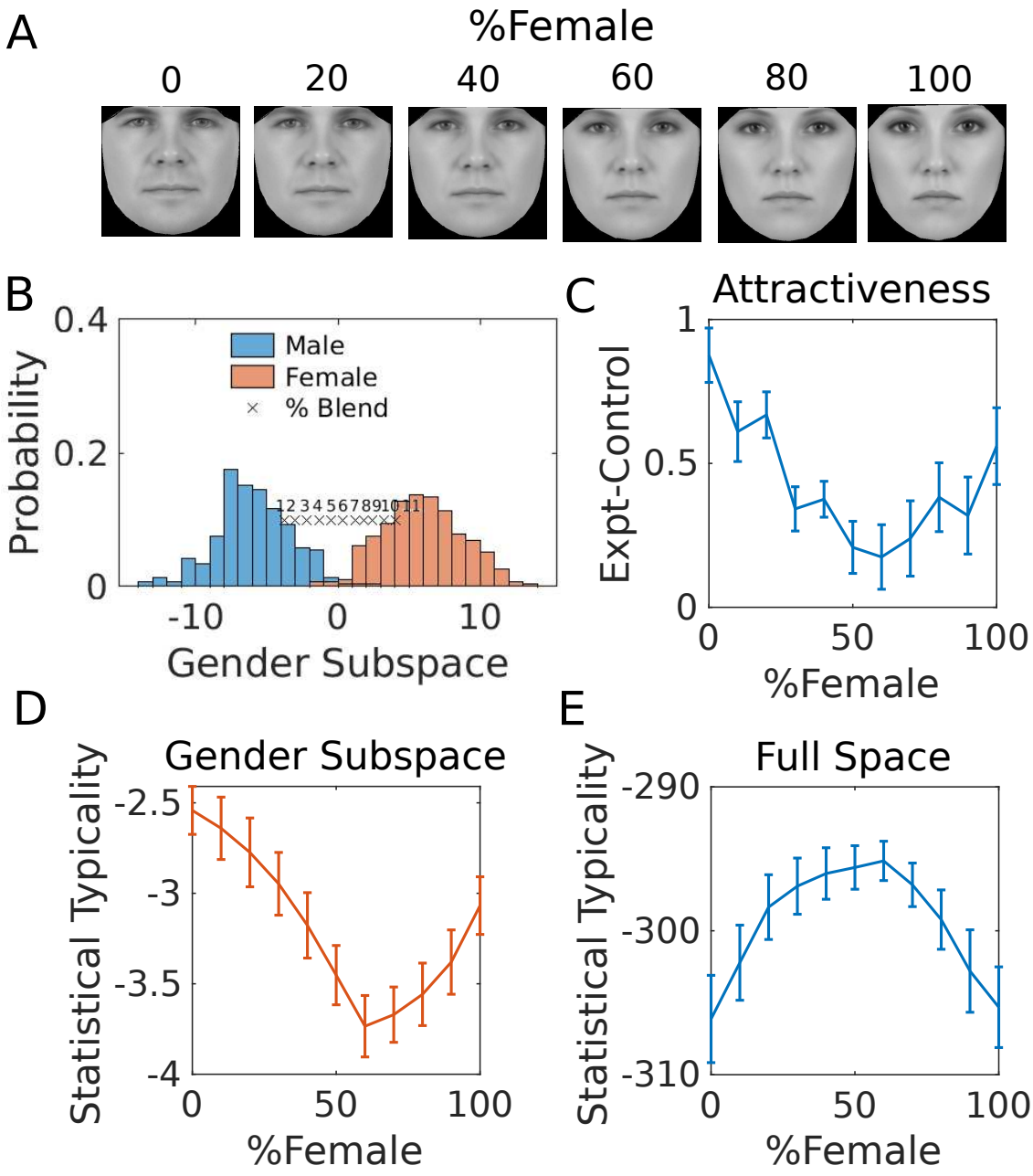

Figure 1.7. UiA of bi-gender blends induced by gender categorization. A. Example stimuli used in [78]: blends of varying proportions of male and female "parent" faces. B. The empirical distribution of male and female faces [60] projected into the gender-informative subspace is a mixture of two approximately normal distributions. X: mean location of actual experimental stimuli [78] for each % of blend. C. *Data*: attractiveness rating in experimental condition minus control condition, as a function of % blend [78]. D. Model-predicted statistical typicality assessed within the gender-discriminating subspace, for the same faces as in C [78]. Model-predicted statistical typicality assessed in the full (original) space, for the same faces as in C [78]. Error bars in C-E: s.e.m over all stimuli used in the experiment for each % blend.

1.3 Discussion

In this paper, we proposed a statistically grounded account of human "liking" of high-dimensional objects, which, in the case of faces, manifests itself as positive social evaluation across multiple traits. We showed that human perception of attractiveness and other positive traits of a face image depends on its statistical typicality, defined as its log likelihood (LL) relative to an internal representation of the face distribution. This hypothesis is motivated by statistical and information-theoretic arguments that a good or efficient representation should maximize the average LL (or equivalently, minimize the average code lengths or cross entropy) of observed data, and is related to the "efficient coding hypothesis" of neural representation [3]. While our analysis is inherently correlational in nature, some existing findings can be re-interpreted to imply that statistical typicality has a *causal* effect on subjective liking. For example, in a neural phenomenon known as "repetition suppression", repeated presentation of the same stimulus, such as a face image, has been shown to increase predicted likelihood of observing the stimulus [113] and thus lead to a robust decrease in evoked neural response [23, 27, 52, 113]. Relatedly, in the psychological "mere exposure" phenomenon, repeated exposure to a novel stimulus, such as a face image, leads subjects to report greater liking [142] and more positive perceived valence associated with the face [7]. Together, it is clear that empirically increased frequency of a stimulus is sufficient to induce both lower neural representational cost and greater subjective liking. Whether the increase in liking is causally mediated by the decrease in neural coding cost is an important direction of future research.

Additionally, we demonstrated that categorization-induced "Ugliness in Averageness" (UiA) effects [38, 78] can be explained as statistical typicality being dynamically, redefined via attentional modulation [17, 139, 101, 137], as a function of task informational

needs. Specifically, some facial features are particularly informative for discriminating gender (or race). The suggestion is that the observer dynamically enhances the processing of these features (mathematically, by restriction to the relevant featural subspace). The statistical distribution of faces is redefined within this subspace (appearing bimodal), thus leading to systematic reassignment of LL to each face. In particular, the faces that straddle the boundary between two categories (e.g. bi-gender or bi-racial blends) tend to have high LL in the original, full face space, but low LL in the dynamically restricted representation – thus resulting in UiA. We showed that this theory can indeed quantitatively capture the categorization-induced changes in liking on an image-by-image basis. One assumption of this theory is that the brain can dynamically alter its representation of faces in a task-dependent manner. Consistent with this, there is evidence that neural receptive fields for faces are rapidly and dynamically modified by attention and task context [32]. Notably, we expect rapid dynamic modulation of stimulus representation to be primarily applicable in the case of featural dimensions that are ecologically relevant (such as those discriminating race and gender). Such dynamic modulation may also be possible for arbitrary featural dimensions but, as logic would suggest and empirical evidence concurs [21, 127], would require extensive additional training. It would be interesting to test in future work whether UiA can also be causally induced by newly learned multi-modal distributions [21, 127], as would be predicted by our model.

Supporting our suggestion that energy allocation in the brain should be efficient and thus proportional to negative LL of the face stimulus, human fMRI studies have shown that energetic expenditure, indexed by BOLD response, across multiple face-responsive areas in the brain is indeed approximately quadratic (negative LL of a multivariate normal distribution is quadratic) [98, 66]. Interestingly, while the initial empirical findings [98, 66, 117] were interpreted to imply that these face-responsive areas explicitly encode

"typicality" or "distinctiveness" of faces, we suggest that any efficient face representation *must* respond in this quadratic manner [103], given the approximately normal distribution of faces that we found here. This is the case whether or not there is an explicit coding of "typicality" or "distinctiveness" in a particular brain area exhibiting quadratic responses to faces. We also note here the important distinction between statistical typicality and subjective typicality. While statistical typicality, of the kind of quadratic signal found in face areas [98, 66], might well contribute to subjective typicality, we have evidence [33] that human judgment of face typicality and memorability (highly related to distinctiveness [5, 18]) both have a strong *linear* component in the face space, just like the four social traits reported here, and thus cannot *only* be driven by statistical typicality or the quadratic signal found in face-responsive areas.

While a detailed neurocomputational theory is outside the scope of this paper, we briefly discuss one plausible, though obviously greatly simplified, neural implementation of our computational-level theory [65]. Various studies have shown that familiar and unfamiliar faces are represented differently in the brain, both in terms of brain regions [53, 27, 71], and coding scheme [83, 9]. In particular, familiar faces appear to be encoded by dedicated feature detectors [83], also known as "grandmother cells," while unfamiliar faces appear to be encoded by a dimensional scheme [9]. Within our framework, one may well ask how the brain encodes a distribution over faces, which is necessary to represent the LL of a new face. One possibility, related to a sampling representation of distributions [129] and the notion of landmark points for manifold learning in machine learning [106], is for the brain to represent the face manifold (and distribution) using a sparsely sampled representation consisting of well-known faces. When a novel face is encountered, the brain could first identify the *closest* previously learned prototype [92, 64, 63, 11] (note the use of "prototype" here differs from social psychology), and then

use a dimensional coding scheme [9] to encode the discrepancy between the retrieved prototype and the novel face. This scheme builds on both prototype- and norm-based face representations [93] and is consistent with the predictive coding hypothesis [84, 143]. Within this framework, a statistically atypical face incurs high coding cost because it tends to be far from the retrieved prototype, and thus the discrepancy will be large and expensive to represent in a dimensional coding scheme [9]. In the UiA-inducing categorization setting, attention enhances the neural response to task-relevant features relative to task-irrelevant features. This has the effect of increasing the overall coding cost of a category-straddling stimulus, since the featural discrepancy (and thus coding cost) between the stimulus and the closest retrieved prototype is highest in the task-relevant dimensions, which are enhanced by attentional modulation. Importantly, this example also illustrates that attentional modulation and statistical typicality alone are sufficient to explain UiA at the neural level, regardless of whether the relationship between statistical typicality and "liking" is due to efficient coding or not.

The above is but one plausible neural coding scheme; however, it illustrates the general notion that the allocation of long-term representational resources (in the form of structural changes such as the formation of a feature detector) are separable from short-term coding cost (in the form of dynamic activity patterns), which is what we hypothesize drives "liking." Well-known faces, which are statistically more typical, are given greater structural representational resources, *in order* to allow it to incur low dynamic coding cost; conversely, statistically atypical faces incur high dynamic coding cost as a consequence of little dedicated structural representation (no dedicated "grandmother" feature detector). While atypical faces in a stable, well-known environment tend to cancel each other out in terms of driving learning, since they are inevitably and symmetrically distributed in the fringes of an approximately normal distribution, a sudden influx of atypical faces (such

as that induced by immigration) could drive systematic representational plasticity so as to reduce average long-term coding cost.

Statistical atypicality (negative LL) is related to the theoretical notion of "unexpected uncertainty", which we earlier proposed to reflect a confluence of unexpected deviations between expectation and observations and signal the need for representational learning [140]. We proposed that unexpected uncertainty, signaled by the neuromodulator norepinephrine, acts in concert with expected uncertainty, signaled by the neuromodulator acetylcholine, to assist the neocortex in learning and maintaining appropriate representations of environmental statistics as well as selecting the appropriate behavioral responses [140]. This theory has received considerable empirical support in the intervening years [69, 70, 68, 67, 80, 88]. Persistently high statistical atypicality of data is precisely the kind of signal that should drive unexpected uncertainty [117]. In the original formulation of expected and unexpected uncertainty [140], we had in mind rather simple kinds of statistical inference and learning such as those related to associative learning; the current work suggests that similar mechanisms may also apply to highly complex stimuli such as faces. Among other implications, this leads to the interesting hypothesis that atypical faces might lead to an elevation of norepinephrine release. Consistent with this, there is evidence that an early and rapid pupil constriction predicts high attractiveness rating for faces, and experimental manipulation of pupil size *causally* affects perceived facial attractiveness in the expected direction [57]. Combined with the finding that phasic increase in pupil size is associated with elevated norepinephrine release in the brain [88], and that phasic pupil dilation enhances learning (both correlationally and causally) in a manner consistent with unexpected uncertainty [70, 136], this suggests that facial atypicality may indeed modulate norepinephrine-mediated control over attractiveness perception and representational learning in a computationally principled manner.

Some readers may be puzzled by our assertion that the negative affect induced by atypical faces leads to increased learning about those faces, instead of encouraging the observer to pay less attention or physically avoid such faces. However, while negative affect can lead to physical avoidance [40], approach/avoidance behavior has many other contributing factors, such as curiosity. In some situations, negative affect does not lead to physical avoidance [40]. In another classical study, it has been found that there is a strong negative correlation ($r = -0.60$) between liking and exploration of face images, such that those faces perceived as least likable are also those that induce the most exploration [39]. Relatedly, novelty has been observed to drive learning and exploration in relation to faces and visual scenes [28, 79].

We do not claim that statistical typicality to be the sole determiner of attractiveness or other social trait evaluations. For example, our model does not explain certain aspects of facial preferences (e.g., sexual dimorphism in face perception [99, 58]), or systematic differences in preference judgment for faces versus other categories of objects [79]. Even in our own analysis (consistent with prior findings [118, 33]), we find a separate linear component, apparently unique to each social trait, for which we do not yet have a principled explanation. Relatedly, we recently found that the liking "function" over the stimulus space may be modified by positive/negative encounters with specific exemplars, which are then extrapolated to the rest of the space depending on the clustering (categorical) structure of the data – in particular, the most statistically typical exemplar of a cluster/category can be most disliked if previous encounters with members of this category have been negative [128]. Nevertheless, statistical typicality already provides a parsimonious and unifying account of several previously proposed causal factors of attractiveness. For example, we found that popular methods for symmetrizing face images also tend to increase statistical typicality. There is also recent work showing that coding

cost at low-level image statistics (e.g. related to small image patches) also decreases attractiveness [89, 41]; notably, coding cost is equivalent to negative LL in an efficient coding scheme, and low-level statistical typicality is a sub-component of general statistical typicality and can be expected to play a partial role in determining attractiveness. From this perspective, it also makes sense that blemishes [55] should decrease attractiveness, as they are statistically irregular and thus expensive to encode.

Our notion of statistical typicality is related to several previously proposed, qualitative explanations of attractiveness perception, in particular in the context of BiA. For example, averageness [54, 90, 110] or distance to the norm (mean face) [41] have been suggested to contribute to facial attractiveness. As discussed above, when the face distribution is approximately multivariate normal and the facial features are uncorrelated, negative LL is essentially squared Mahalanobis distance, which is correlated with Euclidean distance used in previous work [41]. Mahalanobis differs from Euclidean distance by normalizing each coordinate by the standard deviation of the data along that axis. This is a very sensible variation, as it standardizes the distance metric irrespective of the coordinate system being used. Additionally, Mahalanobis distance accounts for correlations between features. These subtle differences between Mahalanobis distance and Euclidean distance may partly explain why in our hands LL is a stronger predictor of attractiveness than in previous work [41]. In addition, our study might find a stronger effect of "averageness" because of the greater statistical power induced by face stimuli that range more broadly across the feature space (in units of standard deviation) than in previous work [41] – although Figure 5 makes it clear that it is not *only* the most extreme stimuli that are driving the effect, but the full range of faces. A bigger difference between our statistical typicality account and an averageness account arises in the case of multi-modal distributions, since LL and averageness are no longer well correlated. In this

vein, UiA effects [38, 78] provide a particularly discriminating case for testing statistical typicality versus averageness in attractiveness perception. Through our analyses, we found that statistical typicality is the more relevant variable. Another prominent notion related to statistical typicality is *prototypicality*. Prototypicality is a measure of "closeness" of a stimulus to the "prototype" [134, 124], implying there are clear, fixed modes in the stimulus distribution. Our account differs in two ways: firstly, it does not assume the "prototypes" to be fixed and pre-defined, but rather task-dependent (e.g. the average male face can be considered a "prototype" in a gender categorization task but not in a race categorization task); secondly, statistical typicality is well-defined even for distributions that have no distinct modes, such as for the fairly flat (close to uniform) distribution of *age*, in which case LL is approximately a constant.

Our statistical typicality account is also related to a rather different explanation of BiA and UiA, known as the fluency account [86, 134, 38, 78]. The fluency account hypothesizes that stimuli, such as category-specific prototypes, may be processed more "fluently" than other stimuli, and human liking of faces and other objects decreases in response to "disfluency" in processing. At a broad level, the fluency account shares with our statistical typicality hypothesis the concept of a human preference for efficiency. However, fluency is not mathematically defined, but empirically measured as response time [38, 78], which surely correlates with coding efficiency, but may also include other factors, such as computational complexity, motor delay or effort, and attention or motivational factors unrelated to coding efficiency. Another close relative to statistical typicality and fluency is "simplicity" [10], in the sense that once a simple explanation/representation is learned, using it to process new data is generally more efficient than using a complex explanation/representation, even if learning or discovering the simple explanation itself may not be trivial. Processing disfluency is also related to effort-based decision making

[132, 43], for it is likely that processing cost and energy expenditure at the biophysical level relate to perceived effort at some level. Future empirical work is needed to clarify whether negative human assessment of atypical face images and other complex stimuli indeed arises from the same root cause and mechanisms as the aversion to cognitively effortful tasks [132, 43]. One potential experiment is to ask subjects to first categorize a face as above or below the median age (say 40), and then rate attractiveness: categorization difficulty/effort is still greatest closest to the decision threshold, but the statistical distribution along the age-relevant dimension is not obviously bimodal and thus does not increase away from the categorization threshold. Thus, a statistical typicality-based coding cost would predict a BiA effect, while a processing disfluency account would predict a UiA effect. In conjunction with this experimental work, future theoretical work is needed to clarify the formal relationship between computational concepts such as statistical typicality, coding efficiency [3], processing efficiency, and more qualitative psychological concepts such as fluency [86], effort [132, 43], and simplicity [10].

In addition to providing a statistically grounded explanation of contextual dependence of human attractiveness judgment, our work also provides some general insight as to how high-dimensional data can be analyzed and stored efficiently in an intelligent system. All elements of the model presented here can be easily generalized to non-face stimuli, such as dogs, birds, butterflies, fish, automobiles, watches, and synthetic dot patterns [36, 37, 134]. Minimizing the average coding cost of statistically distributed data is a desirable goal for any efficient representational system, whether natural or artificial. Another important computational insight is that a system can overcome limitations in its representational and computational capacity by dynamically shifting its featural representation to focus on task-relevant dimensions according to the behavioral context. As such, our work sheds light on one possible functional role played by

attentional selection in the brain: it is one way to dynamically construct subspaces that emphasizes feature dimensions that are most relevant or salient for performing the task at hand [17, 139, 137, 101]. As the literature on attention is broad and confusing in both psychology and neuroscience [141, 137, 101], a productive direction of future research would be to see whether our hypothesized role of attention can help to unify contradicting parts of that larger literature.

Our line of reasoning sheds light on a broader computational understanding of how the brain dynamically encodes and processes complex, high-dimensional data. Faces provide excellent stimuli for investigating such processes, because they are informationally rich, ecologically important, and for which we have a computationally tractable and neurally relevant parametric representation (AAM). We therefore used faces to implement and test concrete ideas about information representation and its contextual modulation in this work. The attractiveness literature also provided a convenient empirical test of our theory. However, we expect that the dynamic representational framework we hypothesize here also affects other cognitive processes, such as working memory, learning, decision-making, and problem-solving, in the sense that all these cognitive processes can benefit from attentional enhancement of task-relevant feature dimensions over irrelevant dimensions. For example, learning to memorize a set of items should be easier if one's attention is focused on the features that make these items easier to organize. Another example is that two stimuli that differ along an attended featural dimension should be easier to discriminate and later recall; and those that differ along an unattended dimension (orthogonal to the attended dimension or dimensions) should be harder to discriminate and later recall. In general, the benefits of cognitive expediency and behavioral accuracy derived from focusing on task-relevant features are broad and multi-faceted, pointing to a promising direction for future research.

1.4 Methods

1.4.1 Formal Model

We assume humans have an internal d-dimensional representation of faces $\mathscr{X}$ [124, 99, 76], in which each face is represented by a vector $\mathbf{x} = (x_1, \ldots, x_d)$ of d real-valued features. We also assume that this face space is endowed with a probability distribution $p_{\mathscr{X}}(\mathbf{x})$, reflecting an internal statistical representation of faces [21]. In the absence of a categorization task (e.g. by gender or race), statistical typicality is defined as the log likelihood (LL) of the face under the full distribution $\log p_{\mathscr{X}}(\mathbf{x})$. When the observer performs a categorization task, such as gender categorization, we assume the brain projects each face $\mathbf{x}$ into the task-relevant subspace $\mathscr{X}\tilde{} \subseteq \mathscr{X}$ to obtain $\tilde{\mathbf{x}}$ (for example, if θ is a basis vector for the task-informative 1-d subspace $\mathscr{X}\tilde{}$, the dot product $\mathbf{x} \cdot \theta$ yields $\tilde{\mathbf{x}}$). It then evaluates the statistical typicality of $\tilde{\mathbf{x}}$ as its log likelihood $\log p_{\mathscr{X}\tilde{}}(\tilde{\mathbf{x}}|c)$ under the class conditional marginal distribution $p_{\mathscr{X}\tilde{}}(\cdot|c)$ that marginalizes over all the dimensions orthogonal to the subspace $\mathscr{X}\tilde{}$, where c is the Bayes-estimated (*a posteriori* most probable) category (e.g. "male" or "female" for gender categorization). This formulation of attentional focusing corresponds to completely eliminating task-irrelevant feature information while leaving attended dimensions untouched, which is likely an exaggeration of brain mechanisms [17, 139], but a "softer" attentional mechanism, that partially suppresses information transmission in the irrelevant dimensions relative to the relevant dimensions, should produce similar results.

1.4.2 AAM

We model faces using the Active Appearance Model (AAM), which we choose for its multiple advantages: neural relevance [9], previous success in modeling human

face space and predicting social judgments compared to alternatives [33, 41, 97], ability to output feature coordinates for novel faces and generate realistic-looking synthetic faces for any coordinate setting, and high transparency and customizability in contrast to commercial software (e.g. FaceGen). We train AAM (see SI) using CFD, a public dataset of 597 demographically balanced face images [60], obtaining for each face 30 *shape features* related to the geometric layout of invariant elements of faces (e.g. eyes, eye brows, nose, mouth, contour of the face) and 60 *texture features* related to pixel variations within and among these elements (results are insensitive to exact number of features, see SI).

1.4.3 Experiment

Participants. 41 (mean age 20.6 years, 24 female) UC San Diego undergraduate students participated in the study in exchange for course credit. Participants gave informed consent before taking part in the study. Approval for the study was given by UC San Diego Human Research Protection Program.

1.4.4 Stimuli

We used AAM to generate a total of 2520 synthetic face images (see Figure S2 for example stimuli). *Procedure.* Participants rated faces "intuitively" on a Likert scale (1-5) for how "attractive", "trustworthy", "dominant", and "positive" (valence) faces appeared. Each image received on average 2.35 ratings per trait.

1.5 Acknowledgements

We thank Jamin Halberstadt for sharing the gender categorization data and helpful discussions.

Chapter 1, in full, is a reprint of the material as it appears in *a*) Ryali CK, Goffin S, Winkielman P, Yu AJ (2020). From Likely to Likable: The Role of Statistical Typicality in Human Social Assessment of Faces. Proceedings of the National Academy of Sciences (PNAS) and *b*) Ryali CK, Yu AJ (2018). Beauty-in-Averageness and its Contextual Modulations: A Bayesian Statistical Account. Advances in Neural Information Processing Systems(NeurIPS). The dissertation author was the primary investigator and author of these papers.

1.A Active Appearance Model

1.A.1 Model Training

We train our own version of the Active Appearance Model (AAM) [22, 13, 123], to construct the face feature space (see Figure 1.8 for a schematic illustration of AAM). The training data are 597 grayscale face images from a publicly available dataset [60], balanced for race and gender, with neutral facial expression taken in the laboratory. First, we use the free software Face++ [1] to automatically label 83 landmarks (e.g. contour points of the mouth, nose, eyes) on each face image. Because the face images may be differentially scaled, shifted, or rotated (in the 2D plane of the image), we use Procrustes (as in [123]) to align the landmarks of each image to the average landmarks (obtained from averaging all raw image landmarks across the dataset). But due to this alignment, the average landmarks have to be recalculated, and Procrustes applied again. After doing so for 50 iterations to ensure convergence, the final landmark coordinates are flattened, centered (subtract the average shape) and subject to PCA to obtain the *shape features*. To obtain *texture features*, the pixel values contained within the landmarks of each image are warped (via triangulation) to the "shape" of the average landmarks, thereby obtaining "shape-normalized" texture, which is flattened, centered (subtract the average texture) and subject to PCA. The coordinates in the face space are the concatenation of the shape features and the texture features. We will sometimes use features and coordinates interchangeably.

1.A.2 Obtaining Coordinates of Novel Faces

Once AAM is trained, to obtain the face space coordinates of a new face image, we follow the same procedure as for getting the features of the original training images

[1] https://www.faceplusplus.com

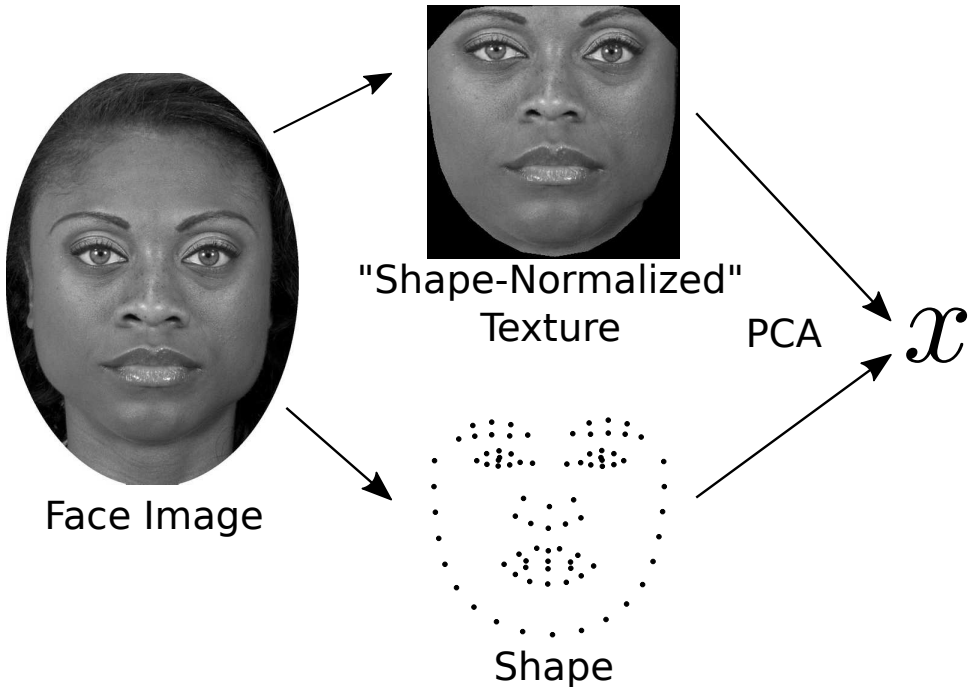

Figure 1.8. AAM-based face representation. A face image is "decomposed" into shape ((x, y) coordinates of Procrustes-aligned landmarks) and shape-normalized texture (grayscale pixel values warped to the "shape" of average aligned landmarks across the dataset). PCA is then conducted over each for dimensionality reduction, and then concatenated to yield a face space.

(without re-training the model).

1.A.3 Generating Synthetic Images

Given a set of features in AAM, we can generate a synthetic image (including stimuli used in the main experiment) by inverting the procedure used to generate the features for a face image. The shape and texture principal components are used to "reconstruct" the shape and shape-normalized texture given the features. The shape-normalized texture is then warped to the generated "shape" to produce a synthetic image.

1.A.4 Creating Face Blends

To create a blend of two (or more) faces, we first obtain the AAM features of the "parent" faces, then average them, and use the averaged features to generate a synthetic image.

1.A.5 Modeling Statistical Typicality (LL)

To model the statistical distribution of faces, we obtain features of CFD faces using the learned AAM representation, and then fit a multivariate Gaussian distribution. Note that we use the CFD dataset twice, once to train the AAM shape and texture features (see Section 1A), and a second time to estimate a density of faces over that feature space.

Statistical typicality (LL) of a stimulus is defined as $\log p(\mathbf{x})$, where $p(\mathbf{x})$ is the CFD-estimated multivariate normal density. Note that because the shape features and texture features are uncorrelated within themselves (but may correlate with each other), LL is a specific quadratic function of the shape features $\{x_i\}$ and the texture features $\{x_j\}$:

$$-\log p(\mathbf{x}) = -\log \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\big(-\frac{1}{2}(\mathbf{x}-\mu)^\intercal \Sigma^{-1}(\mathbf{x}-\mu)\big) \tag{1.1}$$

$$= \sum_i \frac{\tilde{x}_i^2}{\sigma_i^2} + \sum_j \frac{\tilde{x}_j^2}{\sigma_j^2} + 2\sum_{ij} \frac{\tilde{x}_i \tilde{x}_j}{\sigma_{ij}^2} + C, \tag{1.2}$$

where $\tilde{x}_i = x_i - \mu_i$ are centered coordinates, and C is a normalization constant that does not depend on $\mathbf{x}$. Here, μ_i is the mean along dimension i; σ_i^2 and σ_{ij} are the diagonal (variance) and off-diagonal (covariance) terms in Σ. Since we do not know the veridical μ, Σ, we use maximum likelihood estimates. Note that this expression (ignoring C) is the square of Mahalanobis distance; the further away the face is from the mean face (origin), the lower LL is.

Since race and gender correspond to potentially multiple modes, modeling the

distribution of the population faces as a unimodal multivariate Gaussian distribution is an approximation. We quantify the quality of this approximation as follows. We consider two models, a unimodal multivariate Gaussian $p_{\mathrm{G}}(\mathbf{x})$ and a mixture of 2 Gaussians, $p_{\mathrm{MoG}}(\mathbf{x}) = \frac{1}{2}p_{\mathrm{male}}(\mathbf{x}) + \frac{1}{2}p_{\mathrm{female}}(\mathbf{x})$, where the mixture components correspond to genders. Each mixture component is fit by maximum likelihood estimation, but the component label of each face is given by its known gender label in CFD. The Pearson correlation between the LL generated by the two models for each of the CFD faces is very high: $r = 0.95, p < 0.0001$; it is also very high for the synthetic stimuli used in the experiment: $r = 0.985, p < 0.0001$. Moreover, the correlation between attractiveness ratings (provided in CFD) and the LL of $p_{\mathrm{G}}(\mathbf{x})$: $r = 0.16, p < 0.0001$ and $p_{\mathrm{MoG}}(\mathbf{x})$: $r = 0.19, p < 0.0001$ are very similar. In other words, despite the apparent multi-modality due to gender, a single multivariate Gaussian is a very good approximation for the whole data set. This is because gender (and race) induces multi-modality only in a small number of gender- (and race-) discriminating feature dimensions, and genders (or races) are statistically indistinguishable in the great majority of feature dimensions. One way to see this is by using random projections, as noted in the main text, where we see that normality is rejected in only a relatively small fraction (7.6% of the 500 random projections of CFD data). To give an intuitive perspective to this analysis, consider a synthetic mixture of Gaussians that is bi-modal in *every* dimension i, $p(x_i) = \frac{1}{2}\mathcal{N}(x_i|\mu = -1.5, \sigma^2 = 1) + \frac{1}{2}\mathcal{N}(x_i|\mu = +1.5, \sigma^2 = 1)$. We generate 600 samples (300 from each mixture component) and take 500 random projections (a random projection vector $w \in \mathbb{R}^{90}$ generated by i.i.d sampling of each w_i from $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. We find that normality is rejected in 60% of the random projections, implying that a unimodal multivariate Gaussian would not be a good approximation in this case, as expected. This indicates that the small fraction of normality rejection found for the CFD data is not a trivial result, despite central limit

theorem-like arguments related to random projections [20, 87].

1.B Simulation: Beauty-in-Averageness

Similar to the experiment in [54], we create 2-, 4-, 8-, 16-, 32-face blends using male and female faces separately. To create a k-face blend, k distinct faces of the appropriate gender are randomly select from the CFD dataset and their coordinates in the parametric AAM face space are averaged; the averaged coordinates are used to generate a synthetic image.

Similar to the experiment in [38], to produce Figure 2B, a pair of faces are randomly selected from CFD and their coordinates are averaged in varying proportions, from 100% of one face to 100% of the other, and statistical typicality of each blend so produced is evaluated. This procedure is repeated 500 times and the average is plotted in Figure 2B.

1.C Experiment: Social Liking Depends Linearly on Statistical Typicality

1.C.1 Stimuli

Since previous work indicated that BiA depends more strongly on shape than texture features [99], we randomly picked 4 shape PC's and generated face images with 21 pre-set values along these 4 dimensions, $\{-5, -4.5, -4, \ldots, +4, +4.5, +5\}$, in units of CFD sample standard deviation along these axes. Coordinates for the remaining shape features and all texture features were randomly drawn from the multivariate normal density fit to CFD. We generated 10 images per step to produce 210 images for the set of four axes. We then repeated this procedure 10 times, plus two additional sets with 6 randomly chosen, yoked shape features (instead of 4), to produce a total of 2520 images. Example stimuli corresponding to a random axis used in the experiment are shown in

Figure 1.9. Note that even the central face has relatively negative log likelihood because this is a probability density function of high dimensionality, and it takes on statistically sampled values along all orthogonal axes, some of which may be rather atypical. Since orthogonal coordinates are statistically sampled, it is possible for the LL of *individual* face images to be non-monotonic relative to the axis of interest. In this example, the central face has a slightly lower LL than the $+2.5\sigma$ face due to random contributions of orthogonal axes. Relatedly, while some of the synthetic stimuli have rather "extreme" values (e.g. located at 4 or 5 standard deviations), as was also the case in previous studies (e.g. [21]), these synthetic faces may still appear visually not as extreme as one might expect, because only a small number of dimensions (4-6) have "extreme" values, while the remaining (84-86) features tend to have statistically typical values.

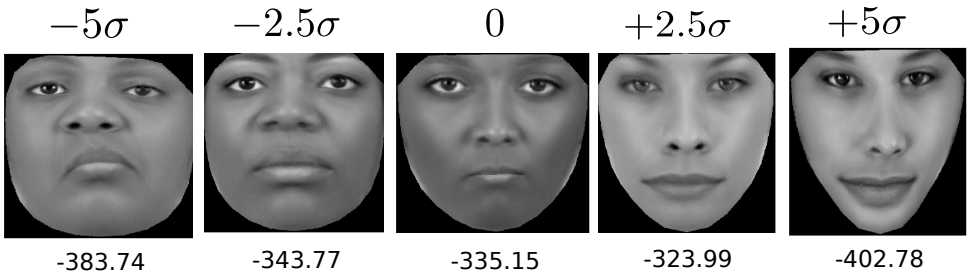

Figure 1.9. Example stimuli used in the experiment. These are five example stimuli corresponding to a particular randomly chosen axis in the AAM face space. The label above each face indicates the feature value of the face along this particular axis, in units of standard deviation of the training data [60] projected along this axis, while taking randomly sampled values on all orthogonal axes. The label below each face indicates its statistical typicality (LL under the multivariate normal distribution fit to CFD [60]).

1.C.2 Rating Standardization and Averaging

We standardize ratings within each participant by removing the mean and dividing by the standard deviation (as was done in [99, 33, 111]) for each trait: attractiveness, trustworthiness, dominance, and valence. We then average the ratings across participants for each image to produce an average rating for a trait on each image.

1.C.3 Model Comparison: BIC Scores

Figure 4B in the main text shows that linear+LL outperforms linear and quadratic models in terms of correlation coefficient between model predicted and subject-reported trait ratings in held-out test face images. Here, we show that linear+LL is also superior in terms of BIC scores (Figure 1.10).

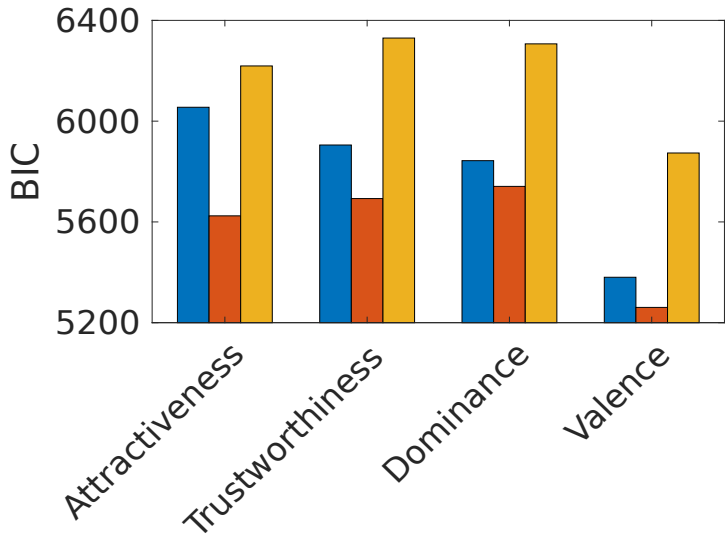

Figure 1.10. Model comparison. BIC for Linear+LL is lowest across traits, consistent with the results using 10-fold CV in Figure 4B.

1.C.4 Individual-Level Correlation Analysis of Social Ratings and LL

Figure 1.11 shows the histogram of Pearson correlation coefficients for subjects who have significant correlations between ratings and statistical typicality: the great majority have positive c.c.

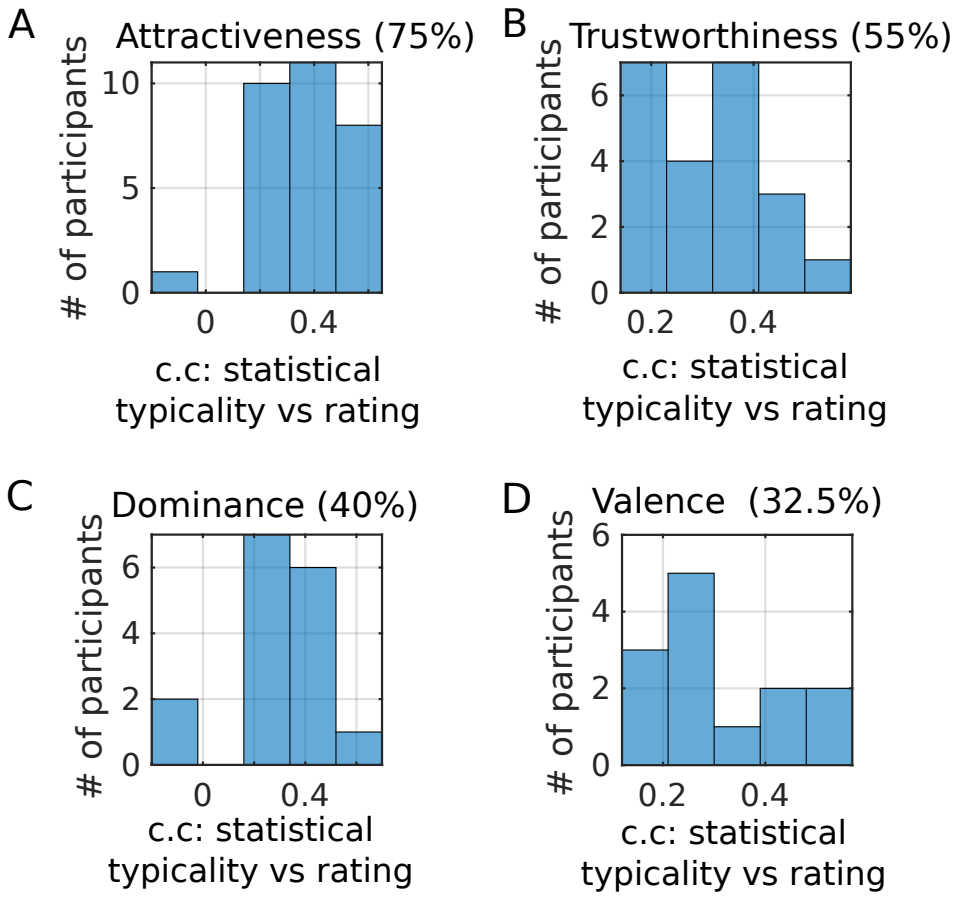

Figure 1.11. Histograms of correlation coefficients between statistical typicality (LL) and standardized ratings of individually significant ($p < 0.05$) participants. Number in parentheses in each panel's title indicates the % of significant subjects for each trait.

1.C.5 Sensitivity of Results to Number of Features

Our findings in the main text are not specific to the choice of 30 shape and 60 texture features used in the analyses. To illustrate this, in this section, we show the same key findings using 25 shape and 25 texture features. First, we report the correlation coefficients between trait ratings and LL of the face stimuli: attractiveness: $r = 0.383, p < 0.0001$, trustworthiness: $r = 0.263, p < 0.0001$, dominance: $r = 0.194, p < 0.0001$, valence: $r = 0.173, p < 0.0001$. As in the main text, residual attractiveness ($r = 0.33, p < 0.0001$) and trustworthiness ($r = 0.178, p < 0.0001$) are significantly correlated

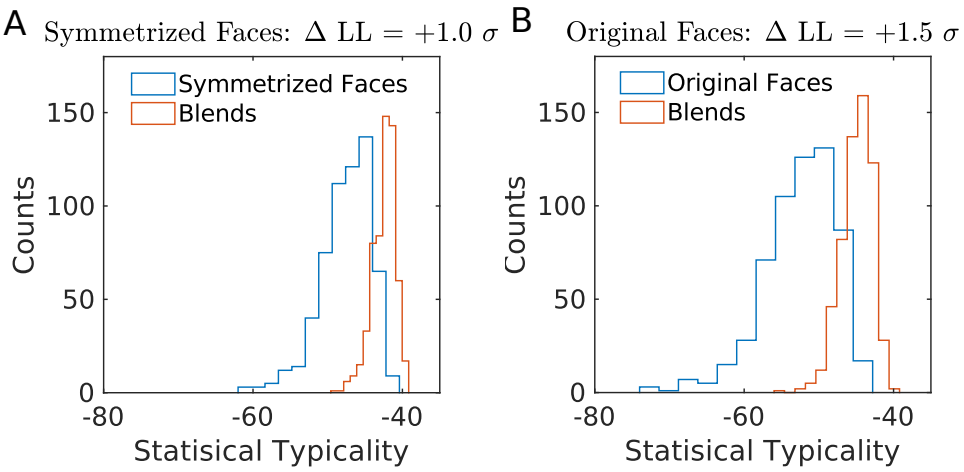

Figure 1.12. Blends of (shape) symmetrized faces still show increase in LL (A), but to a smaller degree than blends of the original (unsymmetrized) faces (B).

with LL.

Model comparison (see Figure 1.14): Linear+LL consistently performs comparably or better than the quadratic model (paired one-sided t-test: $p_{\text{attractiveness}} = 0.1800$, $p_{\text{trustworthiness}} = 0.0140$, $p_{\text{dominance}} = 0.0309$, $p_{\text{valence}} = 0.0279$), as well as the linear model (paired one-sided t-test: $p_{\text{attractiveness}} < 0.0001$, $p_{\text{trustworthiness}} < 0.0001$, $p_{\text{dominance}} < 0.0001$, $p_{\text{valence}} = 0.0002$).

Uniqueness of LTAs (see Table 1.2): the LTA for each trait makes better prediction about the same trait on held-out test data than the LTA for any other trait (paired t-test, $p < 0.0001$), except for the trait "trustworthiness", which is slightly better predicted by the LTA for valence ($r = 0.270$) than for trustworthiness ($r = 0.289$), but this is not statistically significant ($p = 0.191$).

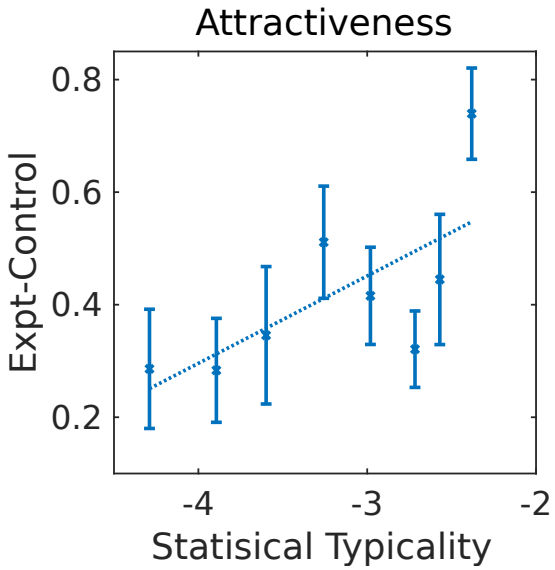

Figure 1.13. Binned scatter plot between (subtractive normalized) attractiveness ratings and statistical typicality in the gender informative subspace for experimental stimuli. Data binning ensures equal number of samples in each bin. Linear regression line (using binned data) is superimposed for visualization. Error bars: s.e.m over stimuli in each bin.

1.D Simulation: Symmetry and Statistical Typicality

1.D.1 Shape Symmetrization

As experimental face symmetrization techniques [46, 100] tend to manipulate only the shape features, Figure 5 in the main text are simulation results based on only symmetrizing the shape features as done in experiments, i.e. for each image, the landmarks and their mirror image are Procrustes aligned to the average landmarks (to remove variation due to translation, scaling, and *roll* or rotation in the image plane) and averaged to produce *symmetrized shape*. The pixel values of the original face image are then warped to the symmetrized shape to produce a shape symmetrized face image. Note that this procedure only changes the shape features and not the texture features of the face.

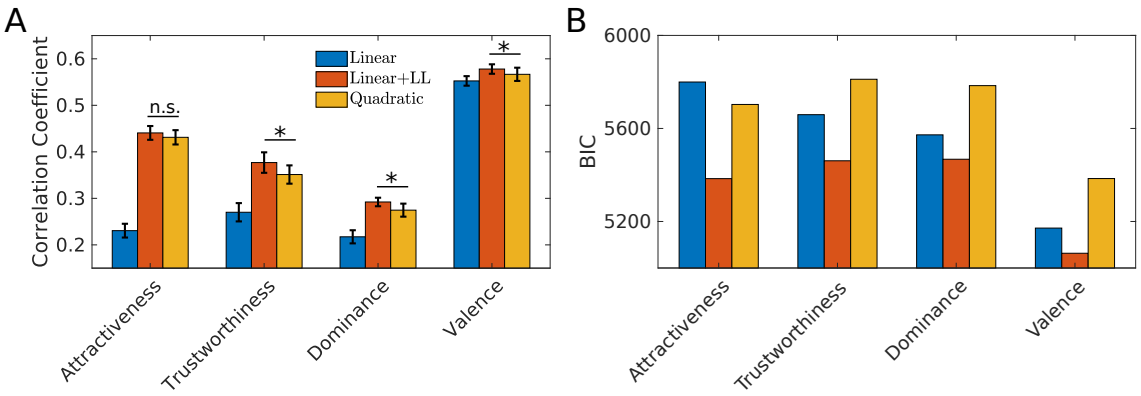

Figure 1.14. Model comparison (Features: 25 shape, 25 texture). A. Correlation coefficient between predicted and actual ratings on held-out data, averaged over 10 folds. Linear+LL consistently performs comparably or better than the quadratic and linear models. Error bars are s.e.m over folds. B. BIC for Linear+LL is lowest. Notation: n.s.=not significant, $\star$: $p < 0.05$, $\star\star$: $p < 0.01$, $\star\star\star$ $p < 0.001$.

1.D.2 Symmetrized Faces still exhibit BiA

It is interesting to note that, blends of pairs of (shape) symmetrized faces still show an increase in LL (Figure 1.12A), though less than the increase in LL for blends of pairs of original (unsymmetrized) faces (Figure 1.12B). This finding is consistent with human behavioral data showing [48] BiA persists after symmetrizing faces, but the effect is smaller than for corresponding unsymmetrized faces.

1.D.3 Symmetrizing Shape and Texture

If in addition to symmetrizing shape features, the texture features are also symmetrized (pixel values of the original and mirror image are warped to the symmetrized shape and averaged), then our model predicts that the increase in statistical typicality (see Figure 1.15) would be even larger (ΔLL=3.4σ), an interesting prediction for future experiments to verify.

Table 1.2. Trait-Specific LTAs (Features: 25 shape, 25 texture). Each trait is generally better predicted by its own LTA than other LTAs, measured by the correlation coefficient between model predictions on ratings of held-out faces and predictions by all LTAs, averaged across the 10 folds of the cross validation.

	Trait Predicted			
LTA	Attractiveness	Trustworthiness	Dominance	Valence
Attractiveness	<u>0.23</u>	0.23	−0.10	0.35
Trustworthiness	0.20	0.27	−0.15	0.50
Dominance	−0.11	−0.17	<u>0.22</u>	−0.27
Valence	0.18	<u>0.29</u>	−0.13	<u>0.55</u>

1.D.4 Additional Information

Statistical typicality in Figure 5 refers to LL evaluated using only the shape features, since texture features remain unchanged. Consequently, ΔLL in Figure 5 is reported in units of standard deviation of LL of the shape features of original faces. We report ΔLL in Figures 5, 1.12, 1.15 in the same units for consistency.

Shape Asymmetry [41] refers to the Euclidean distance between landmarks and their mirror-image after each are Procrustes aligned to the average landmarks.

1.E Ugliness-in-Averageness (UiA)

1.E.1 Gaussian Mixture Modeling of Demographic Subgroups

We model the class (e.g. gender) conditional probability distributions $p(\mathbf{x}|y = \text{male})$ and $p(\mathbf{x}|y = \text{female})$ as multivariate normal distributions $\mathcal{N}(\mathbf{x}|\mu_y, \Sigma_y)$, based on the true gender labels of each face (given in CFD). To estimate the class conditional covariance matrices Σ_y, due to the high dimensionality of the estimation problem (our AAM has 90 dimensions, so each covariance matrix has $O(90^2)$ parameters) and the risk of overfitting, we use cross-validation to select a parameterization of the covariance matrix that can

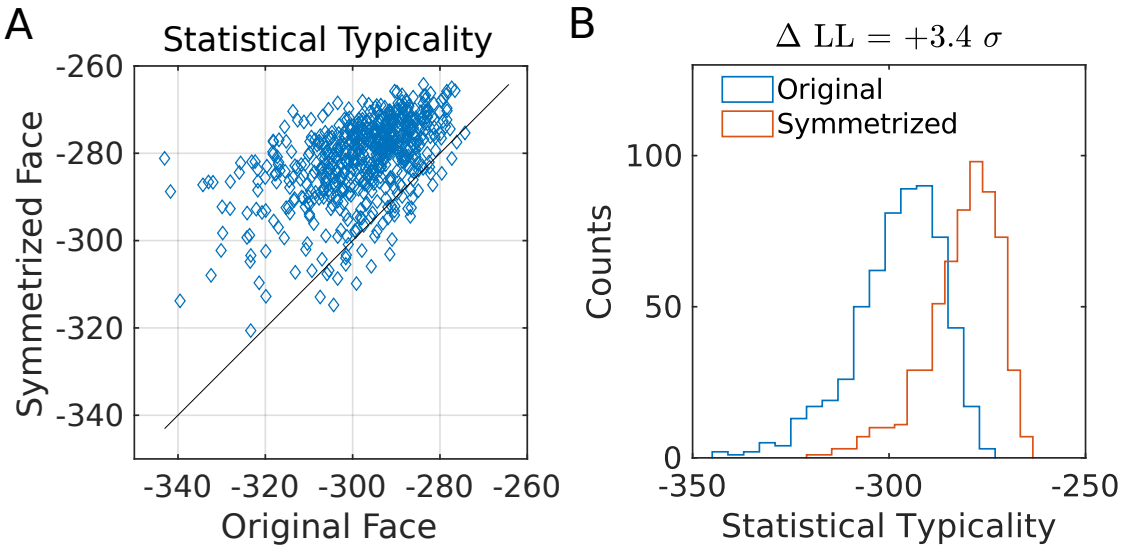

Figure 1.15. Symmetrization of shape *and* texture of a face increases its statistical typicality more than symmetrizing only shape.

best perform gender categorization using MAP estimation based on the resulting model. Concretely, we use 10-fold cross validation to select the structure of the covariance matrices (via the Matlab function *fitcdiscr*) from a) $\hat{\Sigma}_{\text{male}} = \hat{\Sigma}_{\text{female}}$, both diagonal, b) $\hat{\Sigma}_{\text{male}} \neq \hat{\Sigma}_{\text{female}}$, both diagonal, c) $\hat{\Sigma}_{\text{male}} \neq \hat{\Sigma}_{\text{female}}$, non-diagonal d) $\hat{\Sigma}_{\text{male}} = \hat{\Sigma}_{\text{female}}$, non-diagonal, regularized ($\hat{\Sigma}_\gamma = (1-\gamma)\hat{\Sigma} + \gamma \text{diag}(\hat{\Sigma})$), where the regularization parameter $\gamma \in [0,1]$ is found using Bayesian optimization and $\hat{\Sigma}$ is the empirical, pooled covariance matrix. In addition, due to concerns with overfitting, we consider lower-dimensional versions of *a*-*d* (denoted by *a*'-*d*'), by including only the first 30 texture features (in addition to all 30 shape features). Since 10-fold cross-validation shows *d*' to be the best parameterization for both gender (male-female) and race (White-Asian) categorization, we use this covariance structure to evaluate statistical typicality in UiA analyses.

1.E.2 Evaluating Statistical Typicality under Attentional Modulation

Attentional modulation is modeled as a projection into a task-relevant subspace. Given the estimated covariance structure *d*' (see above), the basis vector spanning the 1-*d* subspace that best discriminates e.g. gender, is $\hat{\theta}_{\text{gender}} = \hat{\Sigma}_{\gamma}^{-1}(\hat{\mu}_{\text{female}} - \hat{\mu}_{\text{male}})$, where $\hat{\mu}_{\text{female}}$ and $\hat{\mu}_{\text{male}}$ are means of the male and female distributions in the full space. The projected value of a face image $\mathbf{x}$ in the 1-*d* subspace spanned by $\hat{\theta}_{\text{gender}}$ is just $\tilde{\mathbf{x}}_{\text{gender}} = \mathbf{x} \cdot \hat{\theta}_{\text{gender}}$. The mean of each category i in the projected space is just $\hat{\mu}_i \cdot \hat{\theta}_{\text{gender}}$, and its variance is just the sample variance of the category members projected into this subspace. The prior probability $P(y = \text{male})$ is set to the empirical proportion from CFD [60] (although setting this probability to be uniform, 0.5, produces similar results). Combining the prior with the task-relevant class-conditional likelihood, we can compute the Bayesian *a posteriori* most probable class $y^* = \text{argmax}_y\, p(y|\mathbf{x})$. We then assign statistical typicality in the task-informative (e.g. gender) subspace as $\log p_{\tilde{\mathscr{X}}_{\text{gender}}}(\tilde{\mathbf{x}}_{\text{gender}}|y^*)$.

In order to demonstrate the importance of attentional modulation, we also evaluate statistical typicality in the full face space, i.e. using all features instead of only the gender-informative features. This quantity is evaluated in a manner similar to how statistical typicality in the gender-informative space is evaluated - using the *a posteriori* class conditional, but in the original face space, $\log p_{\mathscr{X}}(\mathbf{x}|y^*)$, where $y^* = \text{argmax}_y\, p(y|\mathbf{x})$. While we use covariance structure *d* to evaluate this quantity in the main text (to keep this quantity on the same scale as measures of statistical typicality in full face space featuring in other analyses), similar behaviour (BiA) is observed if one uses covariance structure *d*' for this computation instead of *d* (see Figure 1.17). Figure 1.18 shows that BiA is also apparent if LL is evaluated not in the full face space but in a randomly projected 1-*d* subspace.

Statistical typicality modeling in the race-informative subspace is analogous.

1.E.3 UiA: Simulation Details

Figure 5B: Similar to the behavioral experiment [38], we generate single-race (Asian-Asian, White-White) blends and mixed-race (Asian-White) blends using female face images. We randomly sample from CFD to generate 100 Asian-Asian, 100 White-White and 100 Asian-White blends. Results similar to Figure 5B may also be obtained using male face images.

Figure 5C;D;E: We randomly sample 60 pairs of Asian and White female faces images from CFD and blend them in increments of 10%. Similar results to Figure 5C;D;E may also be obtained using male face images.

Gender UiA Simulation: Analogous to how mixed-race blends are generated, we randomly sample 60 pairs of male and female face images from CFD and blend them in increments of 10%. Analogous to Figure 6 in the main text, the simulation of gender categorization induced UiA can be seen in Figure 1.16.

1.F Re-analysis of Gender Categorization Data

See [78] for experimental details.

1.F.1 Rating Standardization and Averaging

Ratings are preprocessed by standardizing and averaging in the same manner as previously described in Section 21.C.2.

1.F.2 Evaluating Statistical Typicality

Statistical typicality is evaluated as described in Section 1.E, except actual experimental stimuli [78] are used instead of CFD face images [60].

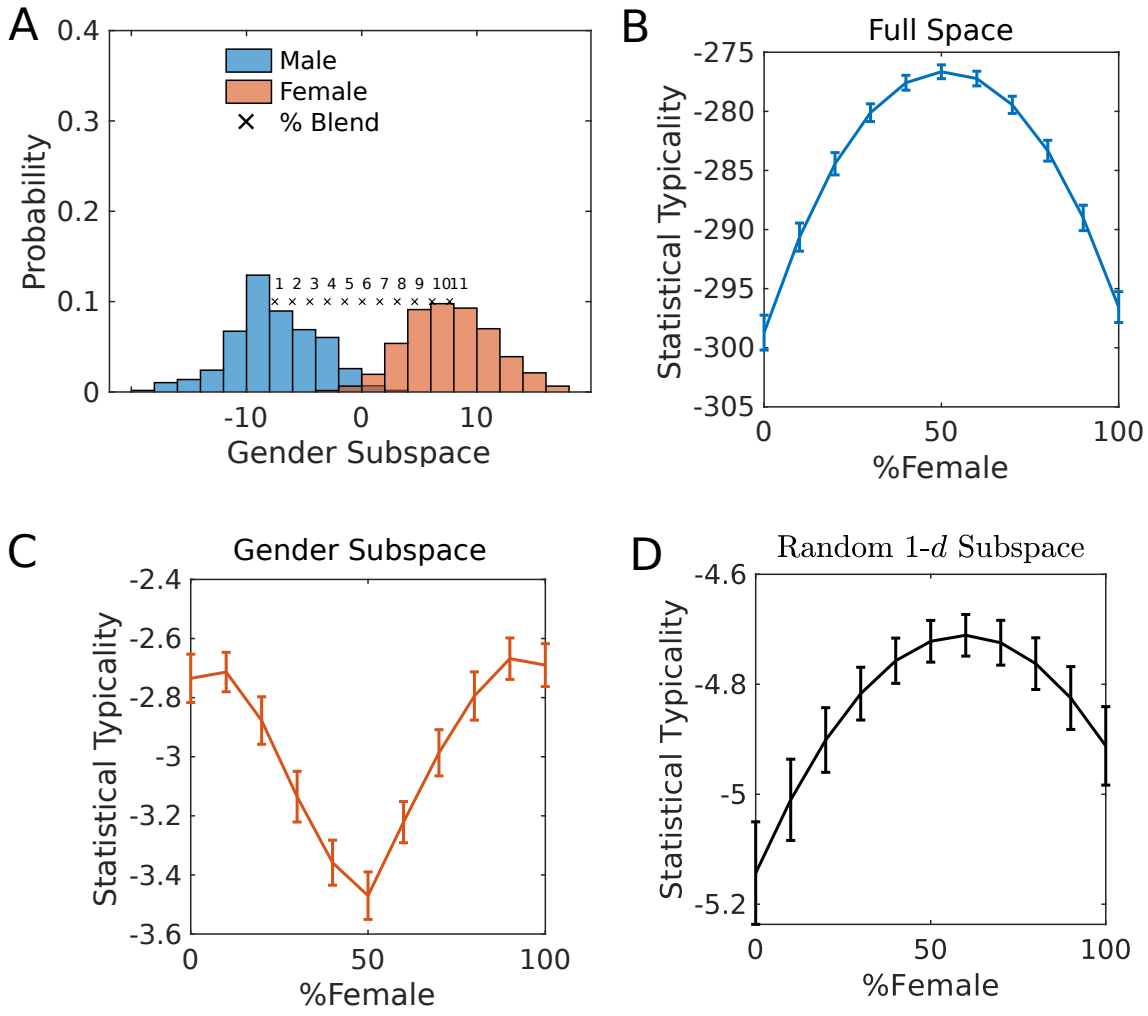

Figure 1.16. Simulation: UiA due to bimodality in the gender-informative subspace. A. The empirical distribution of male and female faces [60] projected into the gender-informative subspace is a mixture of two approximately normal distributions. Crosses (x): mean locations of face images (60 total) for each % of blend, i.e. 1: 100% male and 0% female, …, 11: 0% male and 100% female. B. Model-predicted statistical typicality for face images as a function of % female blend in the original/full face space. C. Model-predicted statistical typicality for face images in the gender-informative subspace. D. Model-predicted statistical typicality for face images as a function of % female blend in a random 1-d subspace. Error bars in B, C, D are s.e.m over simulated samples.

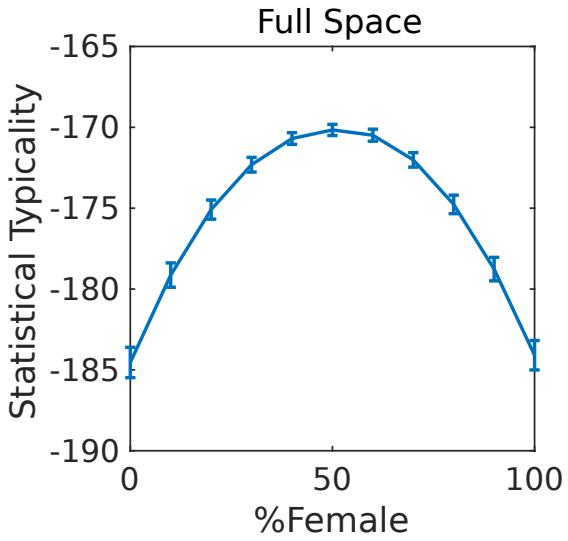

Figure 1.17. Model-predicted statistical typicality for face images as a function of % female blend in the original/full face space using covariance structure d' instead of d also shows BiA, though the scale is different from the otherwise analogous Figure 1.16B. Error bars: s.e.m over simulated samples.

1.G UiA: Familiar Faces

A UiA effect has also been observed [35] when two familiar faces are blended. Here, we discuss how this effect can be captured within the framework proposed. We use a simple abstract model that captures important features of the face distribution. We state the model below.

1.G.1 Generative Model

We assume that humans internally represent each face $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ as generated from a mixture of Gaussians, whereby the components can either correspond to well-known faces $\{f_i\}$ (assume K of these) or demographic subgroups $\{h_r\}$ (assume G

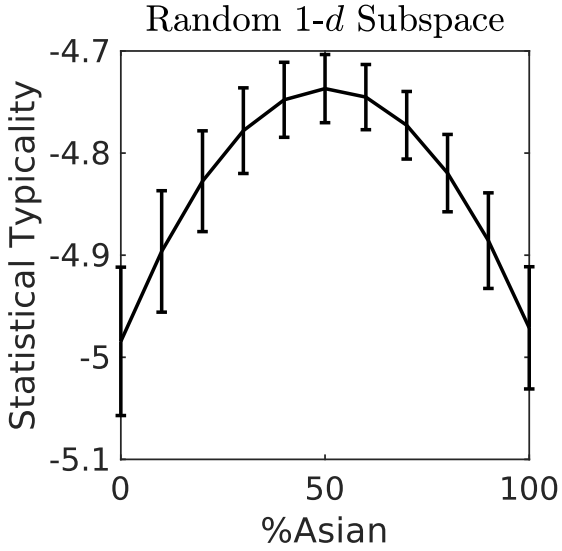

Figure 1.18. Statistical typicality evaluated in a random 1-d subspace shows BiA (as in the full space). Error bars: s.e.m over simulated samples.

of these, e.g. gender, race),

$$X \sim \sum_{k=1}^{|K|} p_k f_k(x) + (1 - \sum_{k=1}^{|K|} p_k) g(x), \tag{1.3}$$

$$g(x) = \sum_{r=1}^{|G|} q_r h_r(x), \tag{1.4}$$

where $h_r(x) = \mathcal{N}(x; \mu_r, \Sigma_r)$, $f_k(x) = \mathcal{N}(x; \mu_k, \Sigma_k)$ and $\sum_{k=1}^{|K|} p_k << 1$ as the number of known faces should be much fewer than unknown faces. We assume that the distributions of the mixture components h_r differ only in a small number of dimensions, $1, \ldots, d_{\text{race}}$ and are identical on the other $d_{\text{other}} := d - d_{\text{race}}$ dimensions. Specifically, we assume $\mu_{r,d_{\text{race}}+1:d} = 0 \in \mathbb{R}^{d_{\text{other}}}$ and

$$\Sigma_r = \begin{bmatrix} \sigma_r^2 \mathbb{1}_{d_{\text{race}} \times d_{\text{race}}} & 0 \\ 0 & \sigma_0^2 \mathbb{1}_{d_{\text{other}} \times d_{\text{other}}} \end{bmatrix}, \tag{1.5}$$

where $\mathbb{1}_{n\times n}$ is an identity matrix of dimensions $n \times n$. For simplicity, we assume $|G| = 2$ and set $\mu_1 = -\mu_2 = \mu$, where $\mu_{1:d_{\text{race}}} = [\mu, \ldots, \mu] \in \mathbb{R}^{d_{\text{race}}}$. We also set the prior/mixture probability distribution q to be uniform.

Approximation

Note that since the statistics of h_r differ only in a small number of dimensions $d_{\text{race}} << d$, the mixture $g(x) = \sum_{r=1}^{|G|} q_r h_r(x)$ is well approximated by $\tilde{g}(x) = \mathcal{N}(x; \mu_0, \Sigma_0)$, where $\mu_0 = 0 \in \mathbb{R}^d$ and $\Sigma_0 = \sigma_0^2 \mathbb{1}_{n\times n}$ and can be assumed to be used to perform inference except when demographic features bear relevance, thus simplifying computations and representation.

Salient feature representation

The mixture components $\{f_k\}$ represent known / recognizable faces, where the variance in each component corresponds to natural variability in a face, such as variations in pose or expressions. For each face k, we assume subjects encode/represent only s distinctive features (relative to the assumed generative distribution) as described in the previous section, denoted by $i_1^k, \ldots, i_s^k$ (the variance along these dimensions is denoted as σ_{sal}^2) and assume the same statistics along other dimensions as $\tilde{g}(x)$, the approximate, assumed generative distribution for a generic, unfamiliar face.

1.G.2 Results

In [35], participants from Netherlands and New-Zealand rated blends of local celebrities (people famous in one country but not the other). Blends of unknown celebrities were rated as more attractive than the "parent" face images (classic BiA), while blends of local celebrities were rated as less attractive relative to the constituent images: a reversal of BiA. An example image (from [35]) depicting a morph of two recognizable faces can

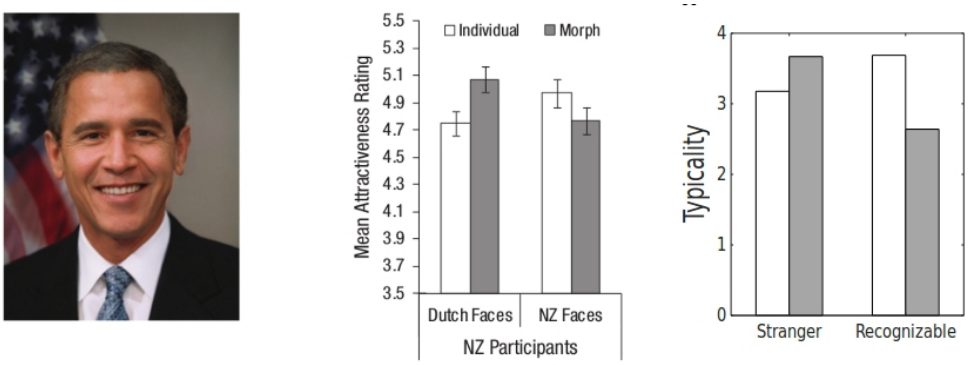

Figure 1.19. UiA in celebrity morphs. (*left*) Illustrative example image (from [35], not actual stimuli) depicting a morph of two recognizable faces (here, Bush and Obama). (*middle*) Blends of recognizable individuals are rated by human subjects as less attractive than individual recognizable faces, while blends of stranger faces are rated as more attractive (adapted from [35]). (*right*) Simulated statistical typicality has similar pattern as data (middle). A constant offset of 6 was added to produce positive values. Simulation parameters: $d = 60$, $d_{\text{race}} = 1$, $s = 2$, $\sigma_0 = 1, \sigma_r = 0.5$ and $\mu = 1$, $|K| = 50$, $\sigma_{\text{sal}} = 0.2$, $\sum_{k=1}^{|K|} p_k = 0.05$, all simulations in 2-d subspace, corresponding to a random subspace or a distinctive feature subspace.

be seen in Figure 1.19 (left), while Figure 1.19 (middle) shows BiA and its reversal in data from the study. Low statistical typicality of the blend in the *distinctive feature subspace* (here 1-d) results in UiA. Simulations qualitatively capture this effect in Figure 1.19 (right).

Chapter 2

Bringing Computer Vision Representations Closer to Human Psychological Representations

2.1 Introduction

Face processing is essential to human social cognition, whether recognizing individuals, identifying emotional states, or assessing social traits such as attractiveness and trustworthiness. Having a computational account of how humans psychologically represent faces is essential for developing and testing scientific hypotheses about human face processing, and for developing machine learning and artificial intelligence systems that either socially interact with humans (e.g. social robots) or mediate social interactions among humans (e.g. dating apps and professional network websites)

An implicit assumption in the psychological study of human face processing is the existence of a "face space" [124], a multidimensional vector space consisting of faces whose vector coordinates correspond to perceived facial properties or features, and the distance between faces determines their perceived similarity. Tools like Multidimensional Scaling (MDS) [104] have been commonly used to leverage similarity judgments to map (embed) faces into a common vector space representation; such representations have

been used to infer mental representations so as to examine perceptual categorization of race [62], to examine the differences in representation between adults and children [74], and to show that faces rated more typical are located closer to the origin while distinctive faces are farther from the origin [45]. Despite its broad use [15, 72, 74, 104, 121], MDS suffers from several limitations. Notably, the mapping of faces into this embedding space is *abstract*, making it difficult to interpret the features; it is *non-invertible*, offering no easy way to visualize the face corresponding to an arbitrary point in the space; it is *non-generalizable*, such that novel faces not used in the learning of the embedding itself cannot be later projected into the space; it is impractical for assessing the *true dimensionality* of the psychological face space, since training MDS-type algorithms are extremely data-intense.

Separately, computer vision and machine learning techniques have been used to learn to predict (or even manipulate) human judgment of different face attributes, e.g. memorability [135, 50], trustworthiness, attractiveness, and other social impressions [111, 33, 96, 95]. However, these work typically do not relate the algorithmic representation of faces to the human face representation, in particular making no attempt to relate distance in the latent representation to human-reported dissimilarity between faces.

Here, we adopt a novel approach, by initializing the face vector space using the latent coordinates of faces generated by different computer vision algorithms, then linearly transforming that vector space such that Euclidean distance in that transformed space recovers human-reported pairwise dissimilarity rating as well as possible – we also include a regularization term that explicitly encourages *efficient* representation. The computer vision algorithms we consider include the Active Appearance Model (AAM) [13], VGG16 [107], and an abstract representation obtained through MDS. As we will show, the AAM-based representation not only predicts human similarity judgements on

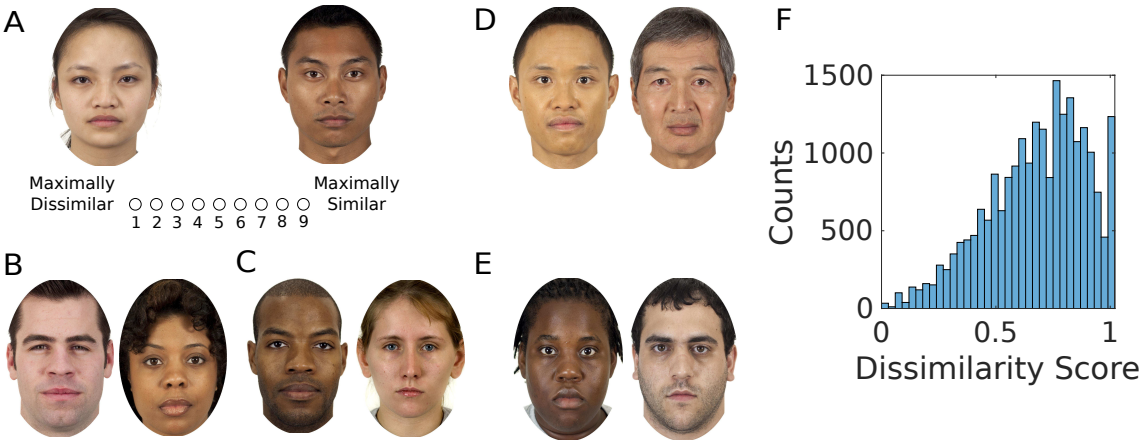

Figure 2.1. A. Schematic of a trial from data collection. B, C: Low-similarity examples. D, E: High-similarity examples. F. Histogram of empirical dissimilarity scores.

held-out data better than the other models as well as other humans who have assessed similarity of the same face pairs, but also performs best in predicting human social trait (e.g. trustworthiness, attractiveness) and affective judgments (e.g. happy, sad, angry).

Using the AAM-based representation, we then investigate several scientific questions, such as how many facial features are actually involved in human perception of how faces differ from one another, whether features that differentiate demographic groups, in particular race and gender, play an especially prominent role in dissimilarity judgments, and whether similarity judgments utilize features that span the entire psychological face space (or whether there are residual features that cannot be excavated using only similarity judgments).

2.2 Results

We collected human similarity judgments on pairs of face images through Amazon Mechanical Turk (restricted to participants based in the US). The data set [60] consists of 595 neutral-expression face images that are gender- and race-balanced (see Methods). Figure 2.1 shows example image pairs with high and low similarity scores. We find

that low-similarity image pairs often differ in race or gender categories, as seen in both low-similarity examples (B, C), while high-similarity pairs can agree on race and gender (D), or not (E). This suggests that human similarity judgments both depend on facial features distinguishing demographic categories and other more subtle structural features.

To model human face representation, we use computer vision models to specify the initial vector space. We first consider AAM [13, 33, 123], which computes "shape features", (x, y) coordinates of landmarks that denote invariant parts of faces such as contours of the eyes, eyebrows, nose, mouths, and "texture features", which are (grayscale) pixel values of each face image warped to have the shape (landmark locations) align with those of the average face in the training dataset. We perform joint principal component analysis (PCA) on the shape and texture features, and retain the first 70 components – as shorthand, we refer to this original AAM space as $\mathscr{X}$. We then linearly transform $\mathscr{X}$ so that Euclidean distances between face images are as close to human dissimilarity scores as possible – formally, this is known as metric learning (see Methods).

A simple way of doing metric learning is to linearly re-scale the importance of each feature (basis vector) in $\mathscr{X}$, i.e. humans may weigh different features differently than the computer vision algorithm. However, it may be that humans actually utilize a different set of features altogether. Formally, we enrich our model by allowing the possibility that psychologically relevant features (basis vectors) are linear transformations of the machine vision features (basis vectors), equivalent to first *rotating* the original feature axes, followed by *rescaling* according to psychological importance in similarity judgment – we denote this linear transformation $\mathbf{W}$.

Additionally, we consider the possibility that humans are *efficient* in the number of features used to represent faces, which we implement through a *regularization* term in the objective function, by explicitly suppressing the number of basis vectors that significantly

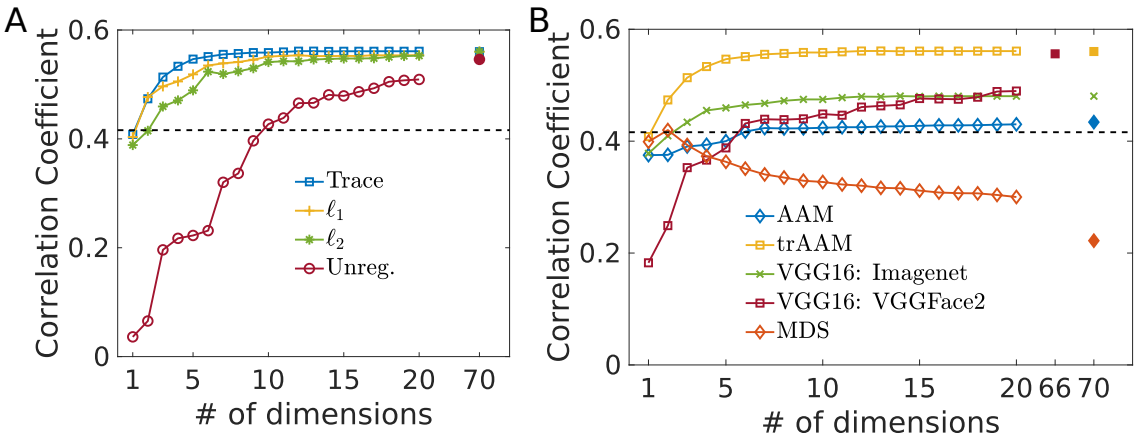

Figure 2.2. A. Effect of regularization on AAM representation. B. Evaluation of various representations; here VGG16 representations correspond to their trace regularized transformed representations. A, B evaluated on validation data (train:validation:test=8:1:1).

contribute to perceptual dissimilarity. Specifically, we penalize the trace of $\mathbf{W}$, or the sum of the squared values of the scaling factors (see Methods). In addition, we also consider two more common forms of regularization, based on penalizing the element-wise ℓ_1 and ℓ_2 norms of the transformation matrix (see Methods), which have the undesirable effect of penalizing not only the scaling factors but the amount of rotation allowed before scaling, and not being especially effective at penalizing the scaling factors.

To compare how well different models can capture/predict human similarity perception, we compute the correlation coefficient (c.c.) between model predicted ratings and human dissimilarity scores on held-out face pairs. As a baseline comparison, the average c.c. between one rater's rating of an image pair and the average rating of the remaining participants on the same image is 0.416. The original AAM representation captures human similarity judgment reasonably well ($r_{\text{test}} = 0.43$), and is significantly improved by the linear transformation without regularization ($r_{\text{test}} = 0.532$). Further prediction improvement is obtained via all three forms of regularization ($r_{\text{test}} = 0.543$ in all cases) on $\mathbf{W}$, all of which prevent overfitting to training data.

In addition to AAM, we also use deep neural networks to initialize the face space

(see Methods). We use VGG16 [107] trained on ImageNet (general object categorization), the best known deep neural network representation for supporting a linear model of human social trait judgement of faces [111]; we also include VGG16 trained on VGGFace2 (face recognition) [6]. Both of these neural networks achieve much worse performance (untransformed: $r_{test}^{\text{VGG16: Imagenet}} = 0.1, r_{test}^{\text{VGG16: VGGFace2}} = 0.31$; transformed: $r_{test}^{\text{VGG16: Imagenet}} = 0.46, r_{test}^{\text{VGG16: VGGFace2}} = 0.53$) than transformed AAM, when only a dozen or so features are included, though they are substantially improved from their untransformed representations; asymptotically, VGG16 trained on VGGFace2 does a comparable job to transformed AAM (Figure 2.2B) – it is interesting to note this model cannot efficiently capture similarity judgments even under trace regularization. We also include a version of MDS (see Methods) for comparison. MDS is comparable to human c.c. with two features, though much worse than computer vision-based algorithms, but its performance steadily deteriorates with more features, reflecting data insufficiency in the absence of an image model.

It is notable that the regularized methods do much better than the c.c. between human ratings on the same image. Human c.c. might have been expected to be a cap on performance, but because human ratings both suffer from within-subject noise, and inter-subject inconsistency, as well as other possible violations of a metric space (e.g. violation of the triangle inequality), one person's rating can be a rather poor predictor of how others will rate the similarity of a face pair; our algorithm can outperform this measure on a novel face because it knows where each face "lives" in the face space relative to other faces, and thus extrapolate from neighboring faces' data to estimate the distance between two new data points.

2.2.1 Dimensionality of Human Similarity Judgment Space.

Among the three types of regularization, we anticipate that trace regularization should be particularly effective in finding a small set of features. Figure 2.2A shows that this is indeed the case. Trace-regularized AAM achieves near-asymptotic performance with many fewer features (most important features first, as indexed by the scaling factor in the transformed space) than ℓ_1- and ℓ_2-regularized AAM. Using only the first 8 features achieves nearly as good of dissimilarity prediction performance ($r = 0.557$) as using all features ($r = 0.561$), while using the first 12 features ($r = 0.561$) is indistinguishable from using all features. Due to the overall superiority of the trace-regularized AAM method in capturing human similarity judgments, we primarily focus on this model in the remainder of the paper (we also sometimes refer to it simply as transformed AAM).

2.2.2 Race- and Gender-Related Features in Human Similarity Judgment.

Figure 2.3A shows synthetic faces generated along each of the first 8 features of the transformed AAM space (ordered by descending value of their scaling factors). Note that the scaling factor of a dimension is indicative of its perceptual importance – Figure 2.3B shows that the average perceptual dissimilarity projected along each dimension (quantifying the average importance of this dimension relative to the overall dissimilarity score) is monotonically related to the scaling factor. All the features appear to be holistic rather than parts-based, and demographic information such as race and gender is clearly present in the first few coordinates, although other more subtle, structural features are also apparent among these featural dimensions. To assess the importance of race- and gender-related features, we consider the average perceptual dissimilarity between subgroups. We note the average model-predicted dissimilarity score between the average male and female faces (0.50), between black and white faces (0.63), between Asian

and black faces (0.57), between Asian and Hispanic faces (0.50), and between Asian and white faces (0.57) are all quite substantial, given that the empirical dissimilarity scores are normalized to have a maximal value of 1 and a minimal value of 0 (see Figure 2.1F for histogram). To quantify this more precisely, we consider the 4D subspace of $\mathscr{X}$ spanned by the axis that differentiates male and female faces (using linear-discriminant analysis, or LDA), and the 3D LDA subspace that best linearly discriminates among the four racial groups. We fit a linear transformation $\mathbf{W}$ within only this subspace – we find that the c.c. between this model-predicted dissimilarity and human-reported dissimilarity on held-out face pairs is $r = 0.44$, or 81% of the performance of using the full model. This indicates race- and gender-informative features figure prominently but not exclusively in human dissimilarity judgments. However, we note that this measure may be somewhat inflated, as the trace regularization suppresses the importance of other features that might also be good at differentiating individual faces but do not add much extra value – in the absence of these race- and gender-informative features, those other features may be able to at least partly make up for the lost capacity and thus achieve c.c. much higher than 19% of the full model.

2.2.3 Face Space: Beyond Similarity Judgments

. Implicit in the concept of a similarity-based "face space" is that features important for similarity judgments also support all other kinds of face-related processing [124, 125], such as race and gender categorization, social trait perception, and affective judgments [33]. Using linear modeling (LDA on categorical discrimination and linear regression on continuous predictions), we can compare how well using only the similarity-relevant features (first 8 dimensions of the transformed AAM, denoted as $\mathscr{Z}$) compares to the original AAM space $\mathscr{X}$, in performing other kinds of tasks. For comparison, we also include VGG16 (trained on either ImageNet or VGGFace2), and MDS. We find that $\mathscr{X}$ is

better or comparable to both deep neural nets and MDS on all tasks (Figure 2.4A: social trait perception, Figure 2.4B: race and gender classification, Figure 2.4C: affect judgements). Compared to $\mathscr{X}$, $\mathscr{Z}$ does slightly worse on social trait perception, similarly on race and slightly worse on gender, and considerably worse on all affective judgments except for "surprise." The general tendency of $\mathscr{X}$ doing slightly better than $\mathscr{Z}$ indicates that certain features unimportant for similarity judgment play a significant role in supporting the other face-based tasks, in particular affective judgments. These results suggest that, in general, it is inadequate to use only similarity judgments to reconstruct the psychological face space, if the goal is to study also other aspects of human face processing.

2.3 Methods

2.3.1 Data Collection

We collected human similarity judgments on pairs of face images through Amazon Mechanical Turk. The stimuli were 595 neutral-expression face images from the Chicago Face Database (CFD) [60], comprising 109 (East) Asian (57 female), 197 black (104 female), 108 Hispanic (56 female), and 181 white (90 female) faces. We randomly sampled pairs of images to produce 23,400 unique pairs, which were rated by 682 raters to produce 138,533 ratings in total. Participants rated the similarity of a pair of face images on a Likert scale from 1 (maximally dissimilar) to 9 (maximally similar); image presentation order was randomized, and subjects rated each image pair twice to counter within-subject variability [130, 112]. To identify non-attentive participants, we included a catch question, where subjects had to indicate if two identical images were the same or not.

2.3.2 Participant Inclusion/Exclusion Criteria

86 raters who failed the catch question were excluded. 4 participants who rated far fewer pairs (< 30) than the other participants (> 200 pairs) were excluded. We also excluded (15) participants whose c.c. of ratings versus other raters on the same images were at least two standard deviations below population mean. We also excluded (32) participants whose response entropy was at least two standard deviations below population mean. Included in the analysis are 111,893 ratings from 551 participants on $22,500$ unique pairs of images (comprising 12.73% of the total possible pairs).

2.3.3 Conversion of Similarity to Dissimilarity Measures

To relate similarity ratings to distances in the face space, we first convert similarity into dissimilarity scores. Let $s^r_{(i,j)}$ denote the similarity rating for images i and j from participant r; we convert it to dissimilarity as $d^r_{(i,j)} = 10 - s^r_{(i,j)}$. We then normalize it for each participant r, $\tilde{d}^r_{(i,j)} = \frac{d^r_{(i,j)} - \min_{i,j} d^r_{(i,j)}}{\max_{i,j} d^r_{(i,j)} - \min_{i,j} d^r_{(i,j)}}$. For each image pair (i,j), we average the normalized dissimilarity ratings to produce an average score $\bar{d}_{(i,j)} = \sum_r \tilde{d}^r_{(i,j)}$. In the main text, we simply refer to the average dissimilarity score as *the dissimilarity score.*

2.3.4 Computer Vision Representation: AAM

AAM is a well-established machine vision technique that reconstructs images well, generates realistic synthetic faces [22], and appears to have neural relevance [9]. AAM consists of *shape features*, or the (x,y) coordinates of a set of consistently defined landmarks (e.g. contours of eyes, noise, lips), and *texture features*, or the grayscale pixel values of a warped version of the image after aligning the landmarks to the average landmark locations across the data set. We train AAM using faces from both CFD and 2222 US adult face images from Google Images [1]. We use the free software Face++ [1] to

[1] https://www.faceplusplus.com

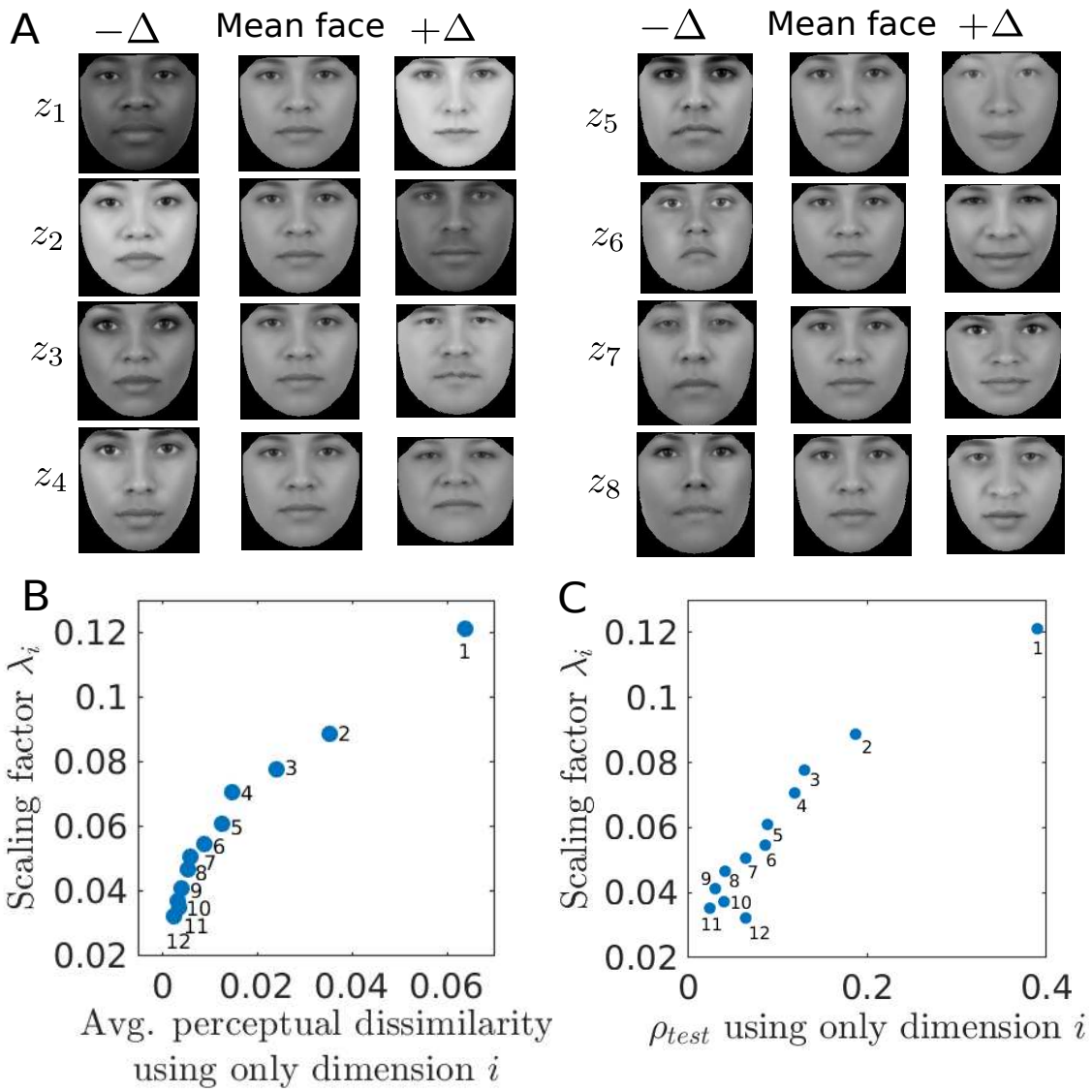

Figure 2.3. Transformed AAM features. A. Synthetic faces along each of the first 8 features (largest eigenvalues of **W**). The stepsize in each direction, Δ, is constant, so that every left/right face compared to the middle face evokes the same amount of perceptual dissimilarity as predicted by the model. B. Scaling factors vs average model predicted perceptual dissimilarities in trAAM along each dimension. C. Scaling factors vs c.c between model predicted and actual dissimilarity scores on test data.

labels 83 landmarks on each face. We apply combined PCA to all the shape and texture features, yielding a 70-dimensional representation that captures 98% of the variance.

2.3.5 Computer Vision Representation: VGG16

VGG16 is a deep Convolution Neural Network (CNN) used for general object recognition [107]. It has been trained using the Imagenet dataset containing 1000 categories of objects, totalling 1.3 million images [94, 19]. Once a face image used in our similarity judgment task is fed into this network, we use the response in the penultimate layer as the image's initial representation. We also use the same architecture trained on VGGFace2 [6] (face recognition). We then perform PCA on extracted features to reduce dimensionality: we retain features capturing 98% of the variance in the CFD dataset (Imagnet-100 PC's, VGGFace2-66 PC's).

2.3.6 Metric Learning

We assume human dissimilarity scores are noisy versions of $f(\mathbf{x}_i, \mathbf{x}_j)$, where $f(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\intercal \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j) + b$, where $\mathbf{W}$ is constrained to be positive semidefinite (PSD; i.e. non-negative eigenvalues) and $b \geq 0$ is a constant offset (b has a fitted value of 0.47 in our main model, trace-regularized AAM). Since $\mathbf{W}$ is PSD, it can be diagonalized as $\mathbf{W} = \mathbf{U}^\intercal \Lambda \mathbf{U}$, where $\mathbf{U}$ is an orthogonal transformation and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_i \geq 0$ are the eigenvalues of $\mathbf{W}$ [2]. Constraining $\mathbf{W}$ to be a diagonal matrix means that the new coordinate system consists of rescaling the original axes, but no rotations are allowed. Allowing $\mathbf{W}$ to be any PSD matrix means the original basis vectors can be rotated and reflected ($\mathbf{U}$ consists of the eigenvectors of $\mathbf{W}$ and specifies the directions of the new basis vectors), and then multiplicatively scaled by the square root of the entries

[2]Note that we're actually modeling the Euclidean distance squared as the the dissimilarity score, as we found this to be empirically better. We may interpret this as modeling a fixed transformation or "link" function of the dissimilarity score (specifically, $(\bar{d}_{(i,j)} - b)^{0.5}$) as the Euclidean distance. We experimented with many other monotonic link functions but did not obtain better results, and will not discuss them here.

of Λ (the eigenvalues of $\mathbf{W}$) to arrive at the new basis vectors. Allowing $\mathbf{W}$ to have 0 as an eigenvalue means that some featural dimensions in the transformed space are allowed to shrink to nothing and thus play no role in perceived dissimilarities.

We then aim to minimize prediction error while regularizing the ℓ_1 or ℓ_2 norm of $\mathbf{W}$. To implement ℓ_1 and ℓ_2 regularization, we minimize the following objective function, denoting $\mathbf{x}_{(i,j)} = (\mathbf{x}_i - \mathbf{x}_j)$ and subject to $\mathbf{W} \succeq 0$, $b \geq 0$,

$$\min_{\mathbf{W},b} \sum_{i,j} (\bar{d}_{(i,j)} - \mathbf{x}_{(i,j)}^{\mathsf{T}} \mathbf{W} \mathbf{x}_{(i,j)} - b)^2 + \alpha \left\| \mathbf{W} \right\|_p$$

where $p = 1$ corresponds to ℓ_1 norm, and $p = 2$ corresponds to ℓ_2 norm. No regularization can be considered a special case ($\alpha = 0$). This is a convex optimization problem, and can be solved via semi-definite programming (we use CVX [30, 29]). We set the value of the regularization coefficient α using line search and evaluation on held-out validation data (choose α that gives the best dissimilarity prediction on the validation set).

To find a small set of interpretable features, we need to suppress the dimensionality of $\mathbf{W}$ (non-zero eigenvalues). ℓ_1 and ℓ_2 regularization are inappropriate because in the former case, both the rotation ($\mathbf{U}$) and the scaling (Λ) components are restricted, while in the latter, the regularization term is not effective at encouraging the eigenvalues to go to zero, $\|\mathbf{W}\|_2 = \sqrt{\mathrm{tr}(\mathbf{W}^\intercal \mathbf{W})} = \sqrt{\mathrm{tr}\big((\mathbf{U}^\intercal \Lambda \mathbf{U})(\mathbf{U}^\intercal \Lambda \mathbf{U})\big)} = \sqrt{\sum_i \lambda_i^2}$. To reduce the number of basis vectors (non-zero eigenvalues), we penalize the sum of the eigenvalues, or trace(Λ) = trace($\mathbf{W}$), resulting in another convex optimization problem (subject to $\mathbf{W} \succeq 0$, $b \geq 0$):

$$\min_{\mathbf{W},b} \sum_{i,j} (\bar{d}_{(i,j)} - \mathbf{x}_{(i,j)}^{\mathsf{T}} \mathbf{W} \mathbf{x}_{(i,j)} - b)^2 + \lambda \ \mathrm{trace}(\mathbf{W}) \ .$$

2.3.7 Multidimensional Scaling

We utilize a version of MDS known as classical MDS [121], which attempts to find coordinates of points in an abstract multidimensional space, such that the inter-point dissimilarities are well-preserved when modeled as Euclidean distances in this space. Consider a graph $\mathscr{G}$ with faces images as nodes, and an edge exists between nodes i and j with length $\bar{d}_{(i,j)}$, if the training dataset contains the dissimilarity score for this pair. Since MDS requires dissimilarities between every pair of images to learn a representation, we estimate the missing distances (edges) as the *shortest* path (sum of edge lengths) between two nodes in $\mathscr{G}$ [102]. Once all pairwise distances have been specified (or estimated), we then run classical MDS to obtain coordinates for all the data points. We also implemented alternative ways to estimate the missing pairwise distances, as well as variants of MDS, but as they achieved poorer similarity prediction performance on held-out data, we will not discuss them further.

2.4 Discussion

In this paper, we presented a novel way of modeling the psychological face space, by first initializing it with a computer vision representation, then linearly transforming it to reproduce human dissimilarity ratings of faces as well as possible. Methodologically, while our broad approach is related to transfer learning [85, 81], we also presented a novel regularization method, that allowed us to make a rather surprising scientific finding: only the 8-12 most important facial features of our model are sufficient to achieve nearly the capacity of the full model to model human face processing, suggesting that the psychological face space may be rather low-dimensional.

By construction, our approach overcomes many of the critical limitations of a common approach in this field [15, 72, 74, 104, 121], namely MDS, by being more interpretable,

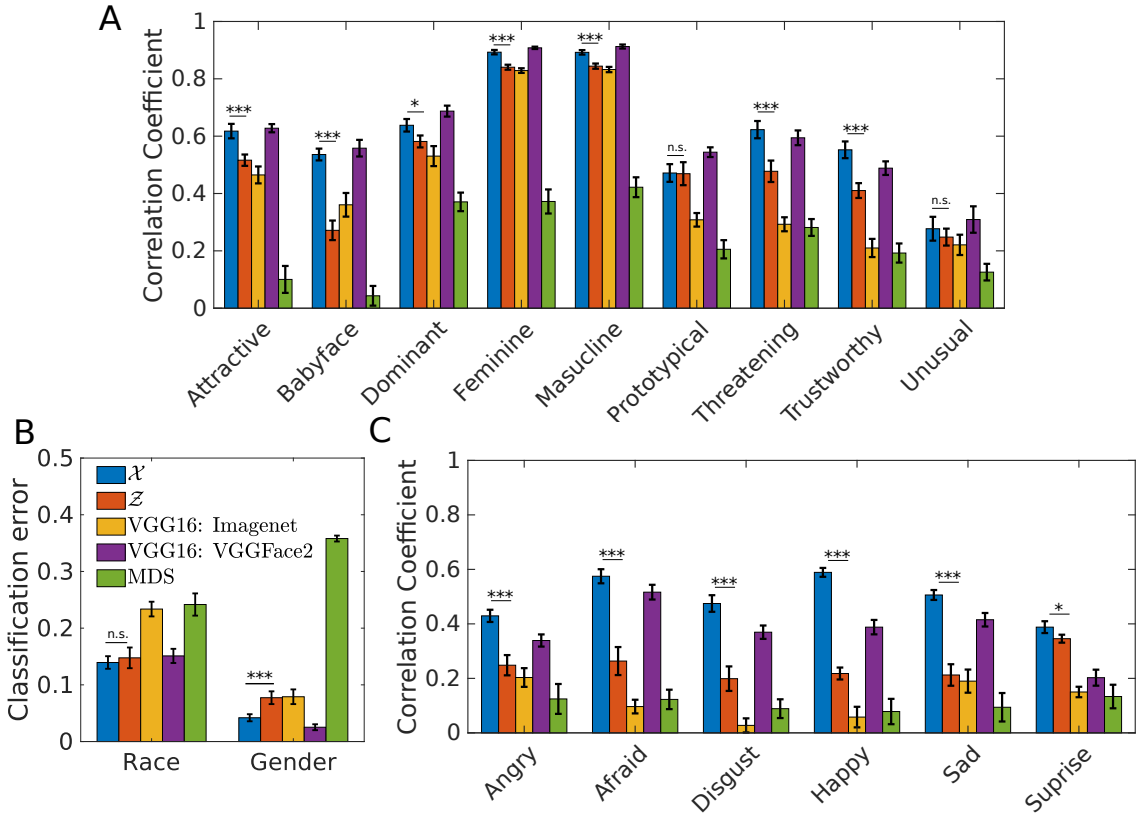

Figure 2.4. A. Social trait prediction. B. Race and gender classification error. C. Emotion ratings prediction. All error bars are SEM over 10-fold CV. Race and gender labels, human ratings of social and emotion traits are from CFD [60]. n.s.: not significant at $\alpha = 0.05, * : p < 0.05, ** : p < 0.01, *** : p < 0.001$; one-sided, paired two-sample t-test.

invertible, generalizable, and data efficient. In addition, we showed that while this method is far better at modeling both dissimilarity judgments and human performance on other face-based tasks (categorizing gender and race, assessing social traits, rating emotional expressions), compared to MDS. However, using only the similarity-relevant features does not work as well as also including the orthogonal features, especially for affective judgments. This scientific finding is at odds with an implicit assumption about human face representation in the psychology literature [124], which, by attempting to reconstructing the full psychological face space using only pairwise similarity judgments, assumes that features important for these judgments are also sufficient for all other face-based tasks.

Another interesting finding is that AAM provides a better initial representation than convolutional deep neural networks trained on both object recognition and face recognition, both for similarity judgments and for other human face-based tasks. We find that VGG16 trained on face recognition (VGGFace2) comes the closest, but is highly inefficient in terms of the number of features it needs to capture similarity judgments (despite having the same trace regularization applied to both). An interesting line of future research would be to consider various unsupervised learning variants of deep neural nets, which may not only learn psychologically relevant features, but also incorporate a decoder model that can generate synthetic images to help visualize/interpret the latent feature space. In particular, adopting techniques that explicitly incorporate inductive biases about shape and texture into the architecture seem promising [105, 73].

2.5 Acknowledgements

We thank Vicente Malave for assistance with data collection and Rongmei Lin for assistance with analysis.

Chapter 2, in full, is a reprint of the material as it appears in Ryali CK, Wang X,

Yu AJ (2020). Leveraging Computer Vision Face Representation to Understand Human Face Representation. Proceedings of the Cognitive Science Society Conference (CogSci). The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] Bainbridge, W. A., Isola, P., and Oliva, A. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General 142*, 4 (2013), 1323–1334.

[2] Bar, M., Neta, M., and Linz, H. Very first impressions. *Emotion 6*, 2 (2006), 269–278.

[3] Barlow, H. B. Possible principles underlying the transformations of sensory messages. In *Sensory Communication*, W. A. Rosenblith, Ed. M.I.T. Press, 1961, pp. 217–234.

[4] Barron, A., Rissanen, J., and Yu, B. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory 44*, 6 (1998), 2743–2760.

[5] Bruce, V., Burton, A. M., and Dench, N. What's distinctive about a distinctive face? *The Quarterly Journal of Experimental Psychology Section A 47*, 1 (1994), 119–141.

[6] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. VGGFace2: A Dataset for Recognising Faces across Pose and Age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (Xi'an, May 2018), IEEE, pp. 67–74.

[7] Carr, E. W., Brady, T. F., and Winkielman, P. Are you smiling, or have i seen you before? familiarity makes faces look happier. *Psychological Science 28*, 8 (2017), 1087–1102.

[8] Caruana, R., and Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning* (2006), 161–168.

[9] Chang, L., and Tsao, D. Y. The Code for Facial Identity in the Primate Brain. *Cell 169*, 6 (2017), 1013–1028.e14.

[10] Chater, N., and Vitányi, P. Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences 7*, 1 (2003), 19–22.

[11] Chen, S. Y., Ross, B. H., and Murphy, G. L. Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General 143*, 1 (2014), 227–246.

[12] Chiu, Y.-C., Esterman, M., Han, Y., Rosen, H., and Yantis, S. Decoding task-based attentional modulation during face categorization. *Journal of Cognitive Neuroscience 23*, 5 (2011), 1198–1204.

[13] Cootes, T. F., Edwards, G. J., and Taylor, C. J. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence 23*, 6 (2001), 681–685.

[14] Cover, T. M., and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, USA, 2006.

[15] Dailey, M. N., Cottrell, G. W., and Busey, T. A. Facial Memory Is Kernel Density Estimation (Almost). In *Advances in Neural Information Processing Systems 11*, M. J. Kearns, S. A. Solla, and D. A. Cohn, Eds. MIT Press, 1999, pp. 24–30.

[16] Davidenko, N., Vu, C. Q., Heller, N. H., and Collins, J. M. Attending to race (or gender) does not increase race (or gender) aftereffects. *Frontiers in Psychology 7* (2016), 909.

[17] Dayan, P., and Zemel, R. S. Statistical models and sensory attention. In *Proceedings of the International Conference on Artificial Neural Networks* (1999), pp. 1017–1022.

[18] Deffenbacher, K. A., and Johanson, J. The face typicality-recognizability relationship: Encoding or retrieval locus? *Memory & Cognition 28*, 7 (2000), 1173–1182.

[19] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference On* (2009), Ieee, pp. 248–255.

[20] Diaconis, P., and Freedman, D. Asymptotics of graphical projection pursuit. *The Annals of Statistics* (1984), 793–815.

[21] Dotsch, R., Hassin, R. R., and Todorov, A. Statistical learning shapes face evaluation. *Nature Human Behaviour 1*, 1 (2016), 0001.

[22] Edwards, G. J., Cootes, T. F., and Taylor, C. J. Face recognition using active appearance models. In *European Conference on Computer Vision* (1998), Springer, pp. 581–595.

[23] Elliott, R., and Dolan, R. J. Neural response during preference and memory judgments for subliminally presented stimuli: A functional neuroimaging study. *Journal of Neuroscience 18*, 12 (2013), 4697–4704.

[24] Friston, K. A theory of cortical responses. *Philosophical Transactions of the Royal Society B 360*, 1456 (2005), 815–836.

[25] Galton, F. Composite Portraits, Made by Combining Those of Many Different Persons Into a Single Resultant Figure. *The Journal of the Anthropological Institute of Great Britain and Ireland 8* (1879), 132–148.

[26] Gauthier, I., Tarr, M., and Bub, D. *Perceptual Expertise: Bridging Brain and Behavior*. Oxford University Press, 2010.

[27] Gobbini, M. I., and Haxby, J. V. Neural response to the visual familiarity of faces. *Brain Research Bulletin 71*, 1-3 (2006), 76–82.

[28] Gottlieb, J., Oudeyer, P. Y., Lopes, M., and Baranes, A. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in Cognitive Science 17*, 11 (2013), 585–93.

[29] Grant, M., and Boyd, S. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds., Lecture Notes in Control and Information Sciences. Springer-Verlag Limited, 2008, pp. 95–110. http://stanford.edu/~boyd/graph_dcp.html.

[30] Grant, M., and Boyd, S. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, Mar. 2014.

[31] Gratton, C., Sreenivasan, K. K., Silver, M. A., and D'Esposito, M. Attention selectively modifies the representation of individual faces in the human brain. *Journal of Neuroscience 33*, 16 (2013), 6979–6989.

[32] Grill-Spector, K., Weiner, K. S., Gomez, J., Stigliani, A., and Natu, V. S. The functional neuroanatomy of face perception: from brain measurements to deep neural networks. *Interface Focus 8*, 4 (2018).

[33] Guan, J., Ryali, C., and Yu, A. J. Computational modeling of social face perception in humans: Leveraging the active appearance model. *bioRxiv* (2018).

[34] Halberstadt, J. The generality and ultimate origins of the attractiveness of prototypes. *Personality and Social Psychology Review 10*, 2 (2006), 166–183.

[35] Halberstadt, J., Pecher, D., Zeelenberg, R., Ip Wai, L., and Winkielman, P. Two Faces of Attractiveness: Making Beauty in Averageness Appear and Reverse. *Psychological Science 24*, 11 (2013), 2343–2346.

[36] Halberstadt, J., and Rhodes, G. The Attractiveness of Nonface Averages: Implications for an Evolutionary Explanation of the Attractiveness of Average Faces. *Psychological Science 11*, 4 (2000), 285–289.

[37] Halberstadt, J., and Rhodes, G. It's not just average faces that are attractive: Computer-manipulated averageness makes birds, fish, and automobiles attractive. *Psychonomic Bulletin & Review 10*, 1 (2003), 149–156.

[38] Halberstadt, J., and Winkielman, P. Easy on the eyes, or hard to categorize: Classification difficulty decreases the appeal of facial blends. *Journal of Experimental Social Psychology 50* (2014), 175–183.

[39] Harrison, A. A. Response competition, frequency, exploratory behavior, and liking. *Journal of Personality and Social Psychology 9*, 4 (1968), 363–368.

[40] Hendricks, M., and Bootzin, R. Race and sex as stimuli for negative affect and physical avoidance. *Journal of Social Psychology 98*, 1 (1976), 111–120.

[41] Holzleitner, I. J., Lee, A. J., Hahn, A., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O. G. B., DeBruine, L. M., and Jones, B. C. Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology: Human Perception and Performance 45*, 12 (2019), 1589–1595.

[42] Ide, J. S., Shenoy, P., Yu, A. J., and Li, C.-S. R. Bayesian prediction and evaluation in the anterior cingulate cortex. *Journal of Neuroscience 33*, 5 (2013), 2039–2047.

[43] Inzlicht, M., Shenhav, A., and Olivola, C. Y. The effort paradox: Effort is both costly and valued. *Trends in Cognitive Science 22*, 4 (2018), 337–349.

[44] Jessen, S., and Grossmann, T. Neural and Behavioral Evidence for Infants' Sensitivity to the Trustworthiness of Faces. *Journal of Cognitive Neuroscience 28*, 11 (2016), 1728–1736.

[45] Johnston, R. A., Milne, A. B., Williams, C., and Hosie, J. Do distinctive faces come from outer space? An investigation of the status of a multidimensional face-space. *Visual Cognition 4*, 1 (1997), 59–67.

[46] Jones, A. L., and Jaeger, B. Biological Bases of Beauty Revisited: The Effect of Symmetry, Averageness, and Sexual Dimorphism on Female Facial Attractiveness. *Symmetry 11*, 2 (2019), 279.

[47] Jones, B., Little, A., Penton-Voak, I., Tiddeman, B., Burt, D., and Perrett, D. Facial symmetry and judgements of apparent health. *Evolution and Human Behavior 22*, 6 (Nov. 2001), 417–429.

[48] Jones, B. C., DeBruine, L. M., and Little, A. C. The role of symmetry in attraction to average faces. *Perception & Psychophysics 69*, 8 (2007), 1273–1277.

[49] Kaminska, O. K., Magnuski, M., Olszanowski, M., Gola, M., Brzezicka, A., and Winkielman, P. Ambiguous at the second sight: Mixed facial expressions trigger late electrophysiological responses linked to lower social impressions. *Cognitive, Affective, & Behavioral Neuroscience 20* (2020), 441–454.

[50] Khosla, A., Bainbridge, W. A., Torralba, A., and Oliva, A. Modifying the Memorability of Face Photographs. In *2013 IEEE International Conference on Computer Vision* (Sydney, Australia, Dec. 2013), IEEE, pp. 3200–3207.

[51] Komori, M., Kawamura, S., and Ishihara, S. Averageness or symmetry: Which is more important for facial attractiveness? *Acta Psychologica 131*, 2 (2009), 136–42.

[52] Koustaal, W., Wagner, A. D., Rotte, M., Maril, A., Buckner, R. L., and Schacter, D. L. Perceptual specificity in visual object priming: Functional magnetic resonance imaging evidence for a laterality difference in fusiform cortex. *Neuropsychologia 39*, 2 (2013), 184–199.

[53] Landi, S. M., and Freiwald, W. A. Two areas for familiar face recognition in the primate brain. *Science 357*, 6351 (2017), 591–595.

[54] Langlois, J. H., and Roggman, L. A. Attractive faces are only average. *Psychological science 1*, 2 (1990), 115–121.

[55] Langlois, J. H., Roggman, L. A., and Musselman, L. What Is Average and What Is Not Average about Attractive Faces? *Psychological Science 5*, 4 (1994), 214–220.

[56] Levy, W. B., and Baxter, R. A. Energy efficient neural codes. *Neural Computation 8*, 3 (1996), 531–543.

[57] Liao, H.-I., Kashino, M., and Shimojo, S. Transient pupil constriction reflects and affects facial attractiveness. *bioRxiv* (2020).

[58] Little, A. C., DeBruine, L. M., and Jones, B. C. Sex differences in attraction to familiar and unfamiliar opposite-sex faces: Men prefer novelty and women prefer familiarity. *Archives of Sexual Behavior 43*, 5 (2014), 973–981.

[59] Luo, L. *Principles of Neurobiology*. Garland Science, 2015.

[60] Ma, D. S., Correll, J., and Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods 47*, 4 (2015), 1122–1135.

[61] Machens, C. K., Golisch, T., Kolesnikova, O., and Herz, A. V. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron 47*, 3 (2005), 447–456.

[62] MacLin, O. H., Peterson, D. J., Hashman, C., and Flach, N. PsychoPro 2.0: Using multidimensional scaling to examine the perceptual categorization of race. *Behavior Research Methods 41*, 3 (Aug. 2009), 668–674.

[63] Malt, B. C., Ross, B. H., and Murphy, G. L. Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition 21*, 3 (1995), 646–61.

[64] Markman, A. B., and Ross, B. H. Category use and category learning. *Psychological Bulletin 129*, 4 (2003), 592–613.

[65] Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., USA, 1982.

[66] Mattavelli, G., Andrews, T. J., Asghar, A. U., Towler, J. R., and Young, A. W. Response of face-selective brain regions to trustworthiness and gender of faces. *Neuropsychologia 50*, 9 (2012), 2205–2211.

[67] McGuire, J. T., Nassar, M. R., Gold, J. I., and Kable, J. W. Functionally Dissociable Influences on Learning Rate in a Dynamic Environment. *Neuron 84*, 4 (Nov. 2014), 870–881.

[68] Meyniel, F., and Dehaene, S. Brain networks for confidence weighting and hierarchical inference during probabilistic learning. *Proceedings of the National Academy of Sciences 114*, 19 (2017), E3859–E3868.

[69] Meyniel, F., Schlunegger, D., and Dehaene, S. The Sense of Confidence during Probabilistic Learning: A Normative Account. *PLOS Computational Biology 11*, 6 (2015), e1004305.

[70] Nassar, M. R., Wilson, R. C., Heasly, B., and Gold, J. I. An approximately Bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *Journal of Neuroscience 30*, 37 (2010), 12366–12378.

[71] Natu, V., and O'Toole, A. J. The neural processing of familiar and unfamiliar faces: A review and synopsis. *British Journal of Psychology 102*, 4 (2011), 726–747.

[72] Nestor, A., Plaut, D. C., and Behrmann, M. Feature-based face representations and image reconstruction from behavioral and neural data. *Proceedings of the National Academy of Sciences 113*, 2 (Jan. 2016), 416–421.

[73] Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., and Yang, Y.-L. HoloGAN: Unsupervised learning of 3D representations from natural images. *arXiv:1904.01326* (Apr. 2019).

[74] Nishimura, M., Maurer, D., and Gao, X. Exploring children's face-space: A multidimensional scaling analysis of the mental representation of facial identity. *Journal of Experimental Child Psychology 103*, 3 (July 2009), 355–375.

[75] Olivola, C. Y., Funk, F., and Todorov, A. Social attributions from faces bias human choices. *Trends in Cognitive Sciences 18*, 11 (2014), 566–570.

[76] Oosterhof, N. N., and Todorov, A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences 105*, 32 (2008), 11087–11092.

[77] O'Toole, A. J., Price, T., Vetter, T., Bartlett, J. C., and Blanz, V. 3D shape and 2D surface textures of human faces: The role of "averages" in attractiveness and age. *Image and Vision Computing 18*, 1 (1999), 9–19.

[78] Owen, H. E., Halberstadt, J., Carr, E. W., and Winkielman, P. Johnny Depp, Reconsidered: How Category-Relative Processing Fluency Determines the Appeal of Gender Ambiguity. *PLOS ONE 11*, 2 (2016), e0146328.

[79] Park, J., Shimojo, E., and Shimojo, S. Roles of familiarity and novelty in visual preference judgments are segregated across object categories. *Proceedings of the National Academy of Sciences 107*, 33 (2010), 14552–14555.

[80] Payzan-LeNestour, E., Dunne, S., Bossaerts, P., and O'Doherty, J. P. The Neural Representation of Unexpected Uncertainty during Value-Based Decision Making. *Neuron 79*, 1 (2013), 191–201.

[81] Peterson, J. C., Abbott, J. T., and Griffiths, T. L. Adapting Deep Network Features to Capture Psychological Representations. *arXiv:1608.02164 [cs]* (Aug. 2016).

[82] Qian, N., and Zhang, J. Neuronal firing rate as code length: A hypothesis. *Computational Brain & Behavior 3*, 1 (2020), 34–53.

[83] Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. Invariant visual representation by single eneurons in the human brain. *Nature 435* (2005), 1102–7.

[84] Rao, R. P., and Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience 2*, 1 (1999), 79–87.

[85] Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops* (Columbus, OH, USA, June 2014), IEEE, pp. 512–519.

[86] Reber, R., Schwarz, N., and Winkielman, P. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review 8*, 4 (2004), 364–382.

[87] Reeves, G. Conditional central limit theorems for gaussian projections. In *2017 IEEE International Symposium on Information Theory (ISIT)* (2017), pp. 3045–3049.

[88] Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., and Tolias, A. S. Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications 7*, 1 (2016), 13289.

[89] Renoult, J. P., Bovet, J., and Raymond, M. Beauty is in the efficient coding of the beholder. *Royal Society Open Science 3*, 3 (2016), 160027.

[90] Rhodes, G., and Tremewan, T. Averageness, exaggeration, and facial attractiveness. *Psychological science 7*, 2 (1996), 105–110.

[91] Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., and Akamatsu, S. Attractiveness of Facial Averageness and Symmetry in Non-Western Cultures: In Search of Biologically Based Standards of Beauty. *Perception 30*, 5 (2001), 611–625.

[92] Rosch, E., and Mervis, C. B. Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology 7*, 4 (1975), 573–605.

[93] Ross, D. A., Deroche, M., and Palmeri, T. J. Not just the norm: Exemplar-based models also predict face aftereffects. *Psychonomic Bulletin & Review 21*, 1 (2014), 47–70.

[94] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision 115*, 3 (Dec. 2015), 211–252.

[95] Ryali, C., and Yu, A. J. Beauty-in-averageness and its contextual modulations: A bayesian statistical account. In *Advances in Neural Information Processing Systems* (2018), vol. 31.

[96] Ryali, C. K., Goffin, S., Winkielman, P., and Yu, A. J. From likely to likable: The role of statistical typicality in human social assessment of faces. *Proceedings of the National Academy of Sciences 117*, 47 (2020), 29371–29380.

[97] Ryali, C. K., Wang, X., and Yu, A. J. Leveraging computer vision face representation to understand human face representation. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (2020), 1080–1086.

[98] Said, C. P., Dotsch, R., and Todorov, A. The amygdala and ffa track both social and non-social face dimensions. *Neuropsychologia 48*, 12 (2010), 3596–3605.

[99] Said, C. P., and Todorov, A. A statistical model of facial attractiveness. *Psychological Science 22*, 9 (2011), 1183–1190.

[100] Saxton, T. K., DeBruine, L. M., Jones, B. C., Little, A. C., and Craig Roberts, S. A longitudinal study of adolescents' judgments of the attractiveness of facial symmetry, averageness and sexual dimorphism. *Journal of Evolutionary Psychology 9*, 1 (2011), 43–55.

[101] Serences, J. T., and Kastner, S. A multi-level account of selective attention. In *Handbook of Attention*, S. Kastner and A. C. Nobre, Eds. Oxford University Press, 2014, pp. 76–104.

[102] Shang, Y., Ruml, W., Zhang, Y., and Fromherz, M. P. Localization from mere connectivity. In *Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking & Computing* (2003), ACM, pp. 201–212.

[103] Shannon, C. E. A mathematical theory of communication. *Bell system technical journal 27*, 3 (1948), 379–423.

[104] Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika 27*, 2 (June 1962), 125–140.

[105] Shu, Z., Sahasrabudhe, M., Alp Güler, R., Samaras, D., Paragios, N., and Kokkinos, I. Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance. In *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11214. Springer International Publishing, Cham, 2018, pp. 664–680.

[106] Silva, J., Marques, J., and Lemos, J. Selecting landmark points for sparse manifold learning. In *Advances in Neural Information Processing Systems* (2006), pp. 1241–1248.

[107] Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]* (Sept. 2014).

[108] Slater, A., Von der Schulenburg, C., Brown, E., Badenoch, M., Butterworth, G., Parsons, S., and Samuels, C. Newborn infants prefer attractive faces. *Infant Behavior and Development 21*, 2 (1998), 345–354.

[109] Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H. J., and Todorov, A. For Your Local Eyes Only: Culture-Specific Face Typicality Influences Perceptions of Trustworthiness. *Perception 46*, 8 (2017), 914–928.

[110] Sofer, C., Dotsch, R., Wigboldus, D. H. J., and Todorov, A. What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science 26*, 1 (2015), 39–47.

[111] Song, A., Li, L., Atalla, C., and Cottrell, G. Learning to see people like people: Predicting social impressions of faces. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (2017), 1096–1101.

[112] Steegen, S., Dewitte, L., Tuerlinckx, F., and Vanpaemel, W. Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology 5* (2014).

[113] Summerfield, C., Trittschuh, E. H., Monti, J. M., Mesulam, M.-M., and Egner, T. Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience 11* (2008), 1004–1006.

[114] Sun, S., Zhen, S., Fu, Z., Wu, D.-A., Shimojo, S., Adolphs, R., Yu, R., and Wang, S. Decision ambiguity is mediated by a late positive potential originating from cingulate cortex. *NeuroImage 157* (2017), 400–414.

[115] Symons, D. *The Evolution of Human Sexuality*. Oxford University Press, 1979.

[116] Thornhill, R., and Gangestad, S. W. Human facial beauty : Averageness, symmetry, and parasite resistance. *Human Nature 4*, 3 (1993), 237–269.

[117] Todorov, A. The role of amygdala in face perception and evaluation. *Motivation and Emotion 36*, 1 (2012), 16–26.

[118] Todorov, A., Dotsch, R., Wigboldus, D. H., and Said, C. P. Data-driven methods for modeling social perception. *Social and Personality Psychology Compass 5*, 10 (2011), 775–791.

[119] Todorov, A., Olivola, C. Y., Dotsch, R., and Mende-Siedlecki, P. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology 66*, 1 (2015).

[120] Todorov, A., and Oosterhof, N. Modeling Social Perception of Faces [Social Sciences]. *IEEE Signal Processing Magazine 28*, 2 (2011), 117–122.

[121] Torgerson, W. S. Multidimensional scaling of similarity. *Psychometrika 30*, 4 (Dec. 1965), 379–393.

[122] Trujillo, L. T., Jankowitsch, J. M., and Langlois, J. H. Beauty is in the ease of the beholding: A neurophysiological test of the averageness theory of facial attractiveness. *Cognitive, Affective, & Behavioral Neuroscience 14*, 3 (2014), 1061–1076.

[123] Tzimiropoulos, G., and Pantic, M. Optimization problems for fast aam fitting in-the-wild. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 593–600.

[124] Valentine, T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology Section A 43*, 2 (1991), 161–204.

[125] Valentine, T., Lewis, M. B., and Hills, P. J. Face-Space: A Unifying Concept in Face Recognition Research. *Quarterly Journal of Experimental Psychology 69*, 10 (Oct. 2016), 1996–2019.

[126] Valla, J. M., Ceci, S. J., and Williams, W. M. The accuracy of inferences about criminality based on facial appearance. *Journal of Social, Evolutionary, and Cultural Psychology 5*, 1 (2011), 66–91.

[127] Vogel, T., Carr, E. W., Davis, T., and Winkielman, P. Category structure determines the relative attractiveness of global versus local averages. *Journal of Experimental Psychology: Learning, Memory, and Cognition 44*, 2 (2018), 250–267.

[128] Vogel, T., Ingendahl, M. N., and Winkielman, P. The architecture of prototype preferences: Typicality, fluency, and valence. *Journal of Experimental Psychology: General* (2020). In press.

[129] Vul, E., Goodman, N., Grifiths, T. L., and Tenenbaum, J. B. One and done? optimal decisions from very few samples. *Cognitive Science 38* (2014), 599–637.

[130] Vul, E., and Pashler, H. Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science 19*, 7 (2008), 645–647.

[131] Wang, Z., Wei, X.-X., Stocker, A. A., and Lee, D. D. Efficient neural codes under metabolic constraints. *Advances in Neural Information Processing Systems 29* (2016), 4619–4627.

[132] Westbrook, A., and Braver, T. S. Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience 15*, 2 (2015), 395–415.

[133] Willis, J., and Todorov, A. First Impressions: Making Up Your Mind After a 100-Ms Exposure to a Face. *Psychological Science 17*, 7 (2006), 592–598.

[134] Winkielman, P., Halberstadt, J., Fazendeiro, T. A., and Catty, S. Prototypes are attractive because they are easy on the mind. *Psychological science 17*, 9 (2006), 799–806.

[135] Xiao, J., Oliva, A., Torralba, A., and Isola, P. What makes an image memorable? In *CVPR 2011(CVPR)* (June 2011), pp. 145–152.

[136] Yu, A. J. Change is in the eye of the beholder. *Nature Neuroscience 15*, 7 (2012), 933–935.

[137] Yu, A. J. Bayesian models of attention. In *Handbook of Attention*, S. Kastner and A. C. Nobre, Eds. Oxford University Press, Oxford, UK, 2014, pp. 1159–1197.

[138] Yu, A. J., and Dayan, P. Inference, attention, and decision in a Bayesian neural architecture. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, Cambridge, MA, 2005, pp. 1577–1584.

[139] Yu, A. J., and Dayan, P. Inference, attention, and decision in a Bayesian neural architecture. In *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT Press, Cambridge, MA, 2005, pp. 1577–1584.

[140] Yu, A. J., and Dayan, P. Uncertainty, Neuromodulation, and Attention. *Neuron 46*, 4 (2005), 681–692.

[141] Yu, A. J., Dayan, P., and Cohen, J. D. Dynamics of attentional selection under conflict: Toward a rational bayesian account. *Journal of Experimental Psychology: Human Perception and Performance 35*, 3 (2017), 700–717.

[142] Zajonc, R. B. Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology 9*, 2 (1968), 1–27.

[143] Zmarz, P., and Keller, G. B. Mismatch receptive fields in mouse visual cortex. *Neuron 92*, 4 (2016), 766–772.