A Model-Based Investigation of the Biological Origin of Human Social Perception of Faces

Sophia J. Huang and **Chaitanya K. Ryali** and **Jianling Liu** and **Dalin Guo**
Jinyan Guan and **Yvonne Li** and **Angela J. Yu**
University of California San Diego
9500 Gilman Drive La Jolla, CA 92093 USA

Abstract

Humans readily form social impressions of faces at a glance, whether assessing trustworthiness, attractiveness, or dominance. However, little is understood about how such computations are carried out neurally. Here, we leverage a computational model of human face perception to quantify and characterize the extent to which macaque monkey face patch neurons encode information relevant for social trait perception. Specifically, we use a social trait prediction model to estimate the social trait ratings for face stimuli viewed by monkeys during a neural recording experiment. We find that, while the monkey face patch neurons are linearly tuned to facial features different from those used by humans to make social judgments, the subspace spanned by the face patch neurons and the subspace spanned by the facial features supporting human social perception are highly overlapping. This result implies that the information present in the monkey face patch neurons are largely sufficient, after *linear decoding*, to support human social perception, thus shedding light on the biological origin of human social processing of faces.

Keywords: face perception; social perception; representation; neural recording; face modeling

Introduction

Face processing plays a special role in human life, as it underpins social interactions essential for survival and reproductive success (Olivola, Funk, & Todorov, 2014). Psychological studies have shown that humans effortlessly and consistently derive social characteristics (social, demographic, emotional traits) from the appearance of faces of strangers (Willis & Todorov, 2006). However, little is known about how such assessments are represented or computed in the brain. In this work, we leverage a computational model of human face perception (Guan, Ryali, & Yu, 2018) to quantify and characterize the extent to which face patch neurons in the macaque monkey brain (Freiwald & Tsao, 2010) encode information relevant for human social perception of faces.

One challenge for studying the relationship between neural responses and human face perception is that human face images are high dimensional and vary among each other in complex ways. To parameterize the space of human face images, we adapt a popular computer vision algorithm, the Active Appearance Model (AAM) (Cootes, Edwards, & Taylor, 2001; Valentine, 1991). AAM provides a vector space representation of face images with several desirable properties. Firstly, this representation is sufficiently rich such that each face image corresponds to a unique point in this space. Secondly, AAM is capable of generating realistic face images, helping to visualize the features encoded by neurons or group of neurons. Thirdly, recent neural data suggest that face patch neurons encode facial features similar to those in AAM (Chang & Tsao, 2017). Here, we train our own version of the AAM model (Guan et al., 2018) using a publicly available face dataset (Bainbridge, Isola, & Oliva, 2013). This dataset also contains human ratings along 20 social trait dimensions, which we model linearly by regressing the trait ratings against AAM latent features. Similarly, we linearly model the classification of gender and age based on human judgments of these qualities on the same face dataset (Bainbridge et al., 2013).

The neural data we analyze are single cell recordings from the face patch areas of the macaque monkey, recorded while the animals viewed 37 human face images (Freiwald & Tsao, 2010) (the original dataset contained 41 face images, in 4 of which the person's eyes are fully or partially closed – these 4 are excluded from our analyses). The face patch areas of the monkey inferotemporal (IT) cortex have been shown to contain neurons that are highly selective for faces (Freiwald & Tsao, 2010). Although face images used in the monkey experiment have not been rated by human subjects for social traits, we can predict the ratings by projecting the face stimuli to our AAM model, and then use the pre-trained regression models to predict the social ratings (Guan et al., 2018).

In the following, we first define and compute each neuron's Linear Response Axis (LRA), the linear axis within AAM that best captures the tuning selectivity of a neuron. We then characterize the properties of the LRA's both individually and as a population. Finally, we compare the facial features encoded by the neuronal LRA's versus those necessary for human social perception.

Results

Predicting Social Trait Perception

In order to predict human social perception of the faces seen by the monkeys, we utilize a model we recently developed based on the Active Appearance Model (AAM) (Guan et al., 2018). The model obtains a latent vector space representation of face images, consisting of combined principal components of shape and texture features (see Methods). We then use linear regression to model how latent features of a face give rise to trait ratings (20 social traits as in (Bainbridge et al., 2013),

plus the demographic traits gender and age, see Methods). We find that this approach predicts human social ratings on a *novel face* better than other humans' rating on the *same face*; it also achieves comparable performance to the state-of-the-art convolutional deep neural network, but has the advantage of having better interpretability (Guan et al., 2018).

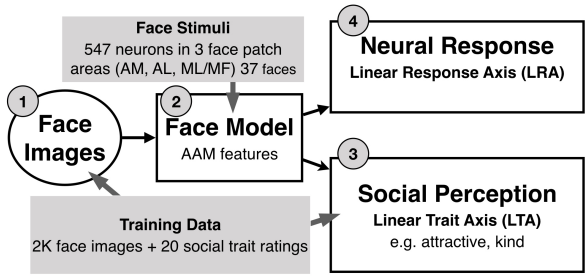

Figure 1: The face model is a vector space representation, whose axes represent the facial features that vary among face images, and the mean of the training dataset sample is centered at the origin (by design). We utilize the publicly available dataset (Bainbridge et al., 2013) (1) to train the AAM face model (2), obtaining 60 facial features. Using the same datasets, we model human social perception (3), and estimate the facial information encoded by each neuron (4).

Here, we can predict the human social ratings of the faces viewed by the monkeys by projecting these face images into the pre-trained AAM model. We first obtain the landmarks of the face stimuli using the free software Face++ (https://www.faceplusplus.com), then projecting them into the pre-trained AAM model (Guan et al., 2018) (Figure 1). Each face stimulus is a point in a 60-dimensional latent space. Figure 2 shows an example face image viewed by monkeys. We then obtain the predicted social perception of each face stimulus using the pre-trained Linear Trait Axes or LTA's (see Methods). The LTA of a trait represents the linear combination of facial features that maximally modulates human perception of this trait (a similar variation in facial feature along any other axis will induce a smaller change in average human perception). For example, the face in Figure 2A is predicted to be slightly more than 1 standard deviation more attractive than the average face (in the training data); Figure 2C shows predicted social ratings a number of traits.

One question we want to answer is how much information related to each social trait is encoded in the neural responses of the monkey face patch neurons. To have sufficient statistical power to assess this, we first need to make sure that the 37 face images span a substantial portion of the predicted trait ratings. This is indeed the case, as can be seen in Figure 3 for "happy" and "attractive." Figure 3A.ii visualizes a pair of face images seen by the monkeys that are predicted by the model to be less (left) or more (right) happy, and another pair (Figure 3B.ii) that is predicted to be less (left) or more (right) attractive. They are consistent with visual intuition.

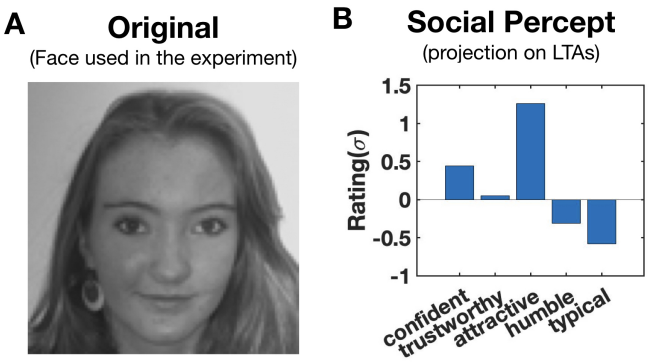

Figure 2: Face representation and social trait estimation. (A) An example face image viewed by the monkeys. (B) 5 examples (out of 20) of predicted social trait ratings for the same face.

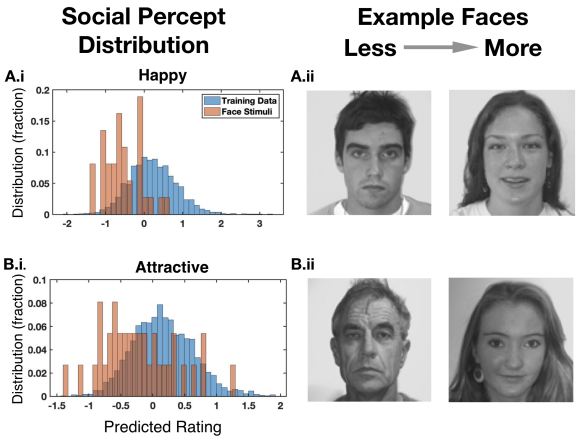

Figure 3: Social trait rating prediction. The histograms (left) of predicted social traits and two face stimuli (right) that are predicted by the model to vary in (A) happiness and (B) attractiveness. Distribution of predicted "happy" rating (A.i) for the 37 face stimuli (red) and training data (blue).

To quantify the information related to human social perception encoded by the macaque face patch neurons, we compute the correlation coefficient between each neuron's mean firing rate for each face (see Methods) and the predicted trait rating of each face. A neuron is deemed to significantly encode a trait if its correlation coefficient has a p-value < 0.05. We find that 19 out of 22 traits are encoded by a significant fraction of the neural population (Figure 4).

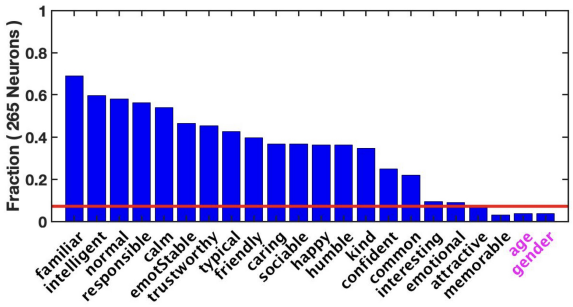

Figure 4: Proportion of neurons significantly encoding various social and demographic traits. A neuron is considered to significantly encode a trait if its MFR has significant correlation ($p < 0.05$, corrected for multiple comparison) with predicted ratings of viewed faces for that trait. The red line indicates the threshold for determining whether a significant (non-zero) fraction of neurons encode a trait at the significance level of $\alpha = 0.05$). This analysis only consists of the 265 neurons whose responses we can statistically reliably model (see subsection on Linear Response Axis)

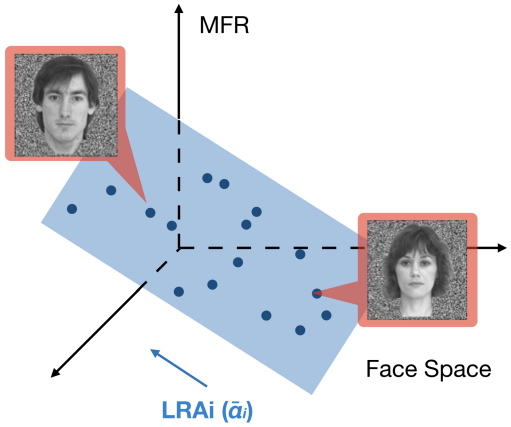

Figure 5: Schematic illustration of Linear response axis (LRA). The blue dots represent MFR of a neuron for different face images. The blue hyper-plane is the best linear fit of the neuron's response to those face stimuli. LRA gives the axis in the face space that yields the largest linear gradient for this neuron's MFR.

Linear Response Axis

To quantify the featural selectivity of each face patch neuron, we first define and compute the Linear Response Axis (LRA) of each neuron (see Figure 5A), which is just the normalized regression coefficient vector. Each LRA is obtained by regressing each neuron's mean firing rate (MFR) against the first $k = 13$ latent features of each image in the AAM space. k is chosen to be 13 in order to maximize the number of neurons whose response we can reliably estimate (i.e. significant correlation between model-predicted MFR and observed MFR on held-out faces, see Methods). For $k = 13$, we find that we can reliably estimate the LRA of 265 neurons – unless otherwise noted, all subsequent LRA-based analyses are performed using only these 265 neurons. The LRA specifies the linear axis that maximally accounts for variations in the neural response. We find that the average neural response along the LRA is not only monotonically increasing, as found in (Chang & Tsao, 2017), but in fact highly linear in this data set; and like in (Chang & Tsao, 2017), the neural response to the principal axis is completely flat. This replicates the finding in (Chang & Tsao, 2017) that monkey face patch neurons encode single axes in the AAM latent feature space.

While Figure 3 quantifies the relationship between social traits and individual neuron's MFR, we are also interested in characterizing the facial features encoded by the neural population as a whole. Naively, we might do so by applying principal component analysis (PCA) to the LRA's. However, the LRA's compose a special sort of data, namely unit-length vectors that lie on a hypersphere. If the LRA's lie in a completely balanced manner (by balanced, we mean that for each LRA, there is an "opposite" LRA that points approximately in the opposite direction, so that the two neurons encode the same AAM axis but have opposite preferences), then PCA would pull out the main directions encoded by neural LRA's; but if they are highly imbalanced, then PCA would instead pull out something like the tangent subspace and yield something uninterpretable. We therefore add an *opposite* LRA to each estimated LRA, to artificially balance the LRA's, and then apply PCA. We find that PC 1 alone explain 48.2% of the total variance among the LRA's, and the first 9 PC's explain 95% of the variance among the LRA's (Figure 6). Relative to the other features, the first PC plays an outsized role in terms of the features that the neurons linearly encode.

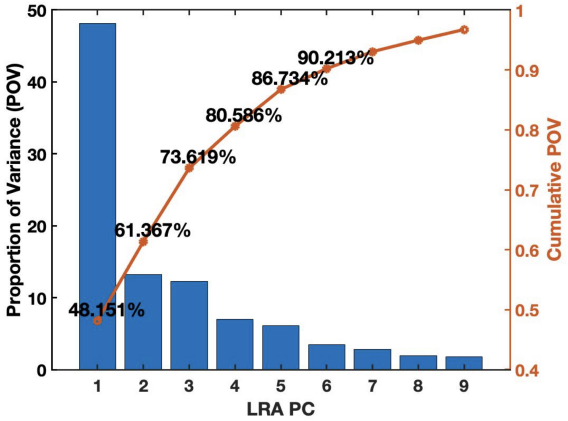

Figure 6: Incremental and cumulative proportion of variance explained by the PCs of neural LRAs. The histogram indicate the explained variance by each LRA PC and the plot for the accumulated explained variance.

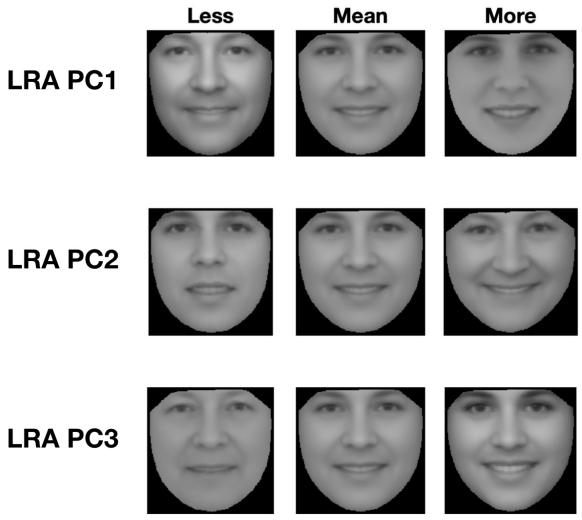

Figure 7: Synthetic face images along the top three LRA PC's. Each row shows how the mean face changes when positive (right) and negative (left) values are added to the mean face along each PC.

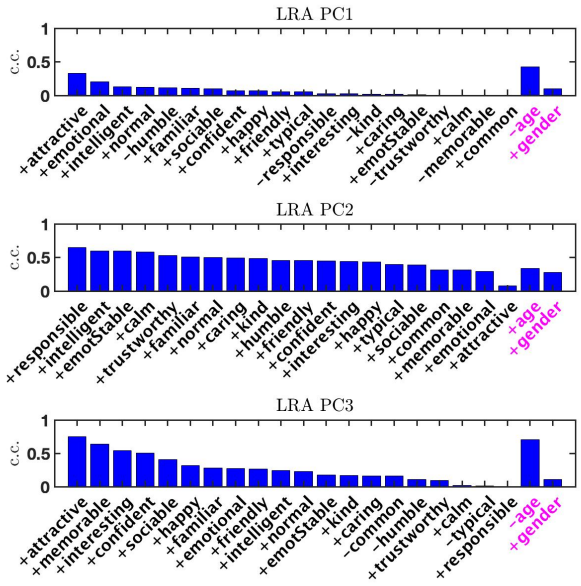

Figure 8: Expected correlation (i.e. dot product) between neural LRA PC's and social trait LTA's. Each row indicate expected c.c. for the various traits for each PC (blue: social trait, magenta: demographic trait). The bars indicate the absolute value of the correlation coefficients while the sign of the correlation is represented by + and - sign in front of the name of traits on x-axis. The traits are ordered in descending order of expected c.c., separately for social and demographic trait.

To get a sense for the primary featural axes that face patch neurons encode, we generate synthetic faces along each of the first three PC's (Figure 7). The first row of Figure 7 shows how the middle face changes as it gains more positive (right face) or more negative (left face) value along the first LRA PC; the next two rows show the same for LRA PC 2 and PC 3, respectively. The faces undergo interesting holistic changes along each of the first three PC's, consisting of some age- and gender-related changes but also other harder-to-verbalize structural changes.

To gain a more quantitative understanding of what the major features the face patch neurons encode as a population, we compute the expected correlation between each LRA PC and social trait (Figure 8), which is just the dot product between each LTA and LRA PC (they are both unit lengths). We find that all three LRA PCs significantly correlate with age. The expected correlation coefficient (dot product) between age LTA and each of LRA PC1, PC2, and PC3 ρ=.48, ρ=-.32, ρ=.67, respectively. In addition, PC1 and PC3 are correlated with attractiveness (PC1: ρ=.35, PC2: ρ=.75), and PC2 correlated with responsible (ρ=.67). This shows that while the neural population as a whole encode features that are highly correlated with those important for human social perception, the most important featural dimension (PC 1) has a poor correlation with any of the human social traits that we considered.

Figure 9 illustrates yet another way to visualize the relationship between neural LRA's and human LTA's. It shows a scatterplot of all the neural LRA's (red), the "pposite LRA's" (gray), and the social LTA's (green) projected into the subspace spanned by the Attractive and Responsible LTA's, the two traits that neurons as a population linearly encoded the most information about. We see that, first of all, that most of the social LTA's are fairly close to unit length within this subspace, indicating that most of them point in a direction very close to this 2D subspace. The neural LRA's show a range of distances from the origin, with the majority lying close to the origin, indicating they primarily point in a direction far away from this 2D plane that is quite important for human social perception. The neurons that do have LRA's pointing close to this subspace (distance close to 1 from the origin) are mostly pointing in the direction of traits such as familiar, intelligent, and normal – the traits that have the greatest number of significant correlation with neurons (Figure 4).

Conversely, we can also visualize all the projected LTA's and LRA's in the subspace spanned by LRA PC 1 and PC 2 (Figure 10). Here, we see that the neural LRA's are highly imbalanced, with most LRA having a positive projection along PC 1. We also see that most human LTA's point in a direction far away from PC 1, but have a fairly large component pointing in the direction of PC 2 (the exception is Attractive, which has the opposite pattern).

Similarly, we can also visualize all the projected LTA's and LRA's in the subspace spanned by LRA PC 1 and PC 3 (Fig-

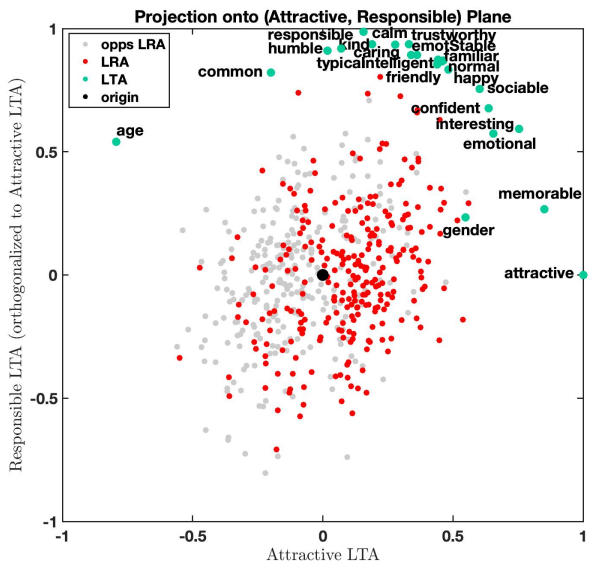

Figure 9: Projection of LRA's and LTA's onto the plane spanned by Attractive and Responsible LTA's. Red dots are projections of LRA's (gray for "opposite" LRA's). Green dots are projections of LTAs'. The label next to each green dot indicates the trait.

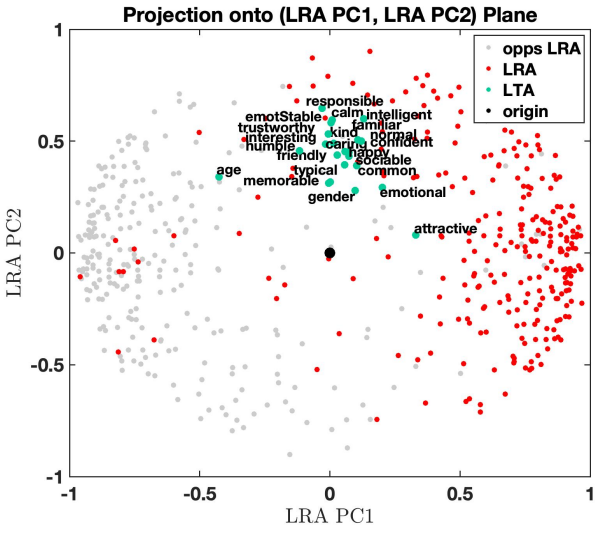

Figure 10: Projection of LRA's and LTA's onto LRA PC1 and LRA PC2. Similar formatting as Figure 9.

ure 10). It is apparent that Attractive and Age have fairly large components pointing along PC 3, along with traits such as memorable, interesting, and confident.

Subspace Comparison

While the previous analyses suggest that there is some overlap in the facial features that are encoded by monkey face patch neurons, and those that matter for human social trait perception, here we quantify their overlap in a different way. We compute how well (model-predicted) human social perception can be computed from the information present in the monkey face patch neurons (via simple linear decoding), and vice versa. As shown in Figure 12, the LTA-predicted ratings of the 22 social traits can be almost perfectly recovered from the LRA-predicted neural response (265 neurons) to face images (R^2 very close to 1); conversely, we find that the LRA-predicted response of all 265 neurons can be perfectly recovered from the LTA-predicted social trait ratings (22 traits), where $R^2 = 1$ in every case. This result suggests that facial featural information present in the macaque face patch areas is largely the same as those necessary for human social perception.

Methods

The Face Model: AAM

The Face model is an instantiation of the Active Appearance Model (Cootes et al., 2001; Guan et al., 2018). Each face image has shape and texture features. The shape features consist of the (x, y) coordinates of a set of landmarks that are consistently defined across faces. The texture features are the pixel values (grayscale) of each face image after warping it to have the same landmark locations as the averaged face. To reduce the dimensionality and remove correlation between shape and texture features, we perform additional Principal Component Analysis (PCA) on shape and texture features and retain the first 60 PC's, resulting in a 60-dimensional AAM feature space. AAM features form the basis of the Face Model that jointly describes the variations of shape and texture of the faces.

Social Trait Perception: Linear Trait Axis (LTA)

The Linear Trait Axis (LTA) $\tilde{\beta}$ for each social trait is computed as the normalized regression coefficients of ratings regressed against AAM features:

$$y = \beta\vec{x} + \varepsilon$$

where y is the standardized ratings for the trait, $\vec{x}$ is the AAM features, and β is the vector of regression coefficients. The linear trait axis (LTA) is defined as

$$\tilde{\beta} = \frac{\beta}{||\beta||}$$

The LTA specifies a direction in the face space that would (linearly) maximally alter the perception of the trait.

Predicted social perception A novel face images can be projected into the trained face model, resulting in a 60-dimensional representation $\vec{x}$. The predicted rating of a face image is then given by

$$\hat{y} = \beta\vec{x}$$

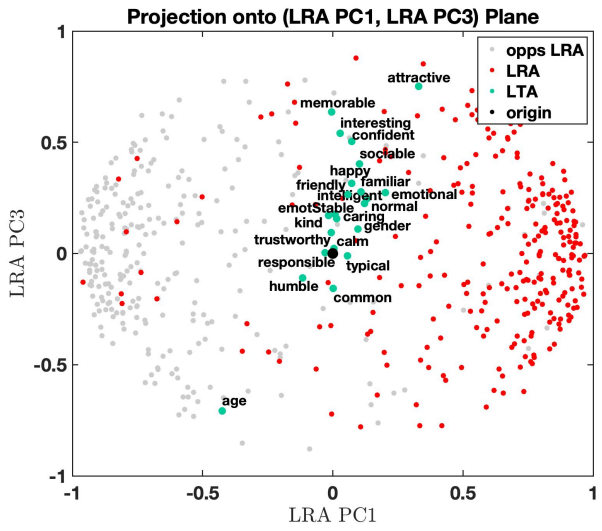

Figure 11: Projection of LRA's and LTA's onto LRA PC 1 and LRA PC 3. Similar formatting as Figure 9.

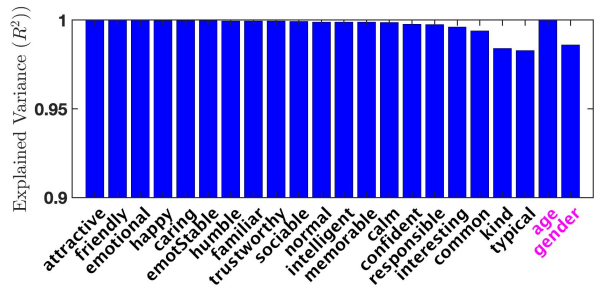

Figure 12: Comparison between LRA subspace and social LTA subspace. The bars (blue: social trait, magenta: demographic trait) represent the amount of variance in LTA explained by LRA, measured in terms of R^2.

with $\vec{x}$ as the 60-dim representation in the face model, and β the regression coefficients for the target trait.

Neuron Encoding - Linear Response Axis (LRA)

Similar to LTA, the Linear Trait Axis (LRA) $\tilde{\alpha}$ for each neuron is the normalized regression coefficients of neurons mean fire rate (MFR) regressed against AAM features.

$$r = \alpha\vec{x} + \varepsilon$$

$$\tilde{\alpha} = \frac{\alpha}{||\alpha||}$$

where r is the neurons MFR, $\vec{x}$ is the AAM features, and α is the vector of regression coefficients. $\tilde{\alpha}$ is the axis in the Face Model that drives maximal (linear) variation of the neurons response. Consistent with existing literature (Chang & Tsao, 2017), we find that MFR (averaged across neurons) increases monotonically along the LRA, and is flat along the principal orthogonal axis (data not shown).

Cross-validation is implemented to evaluate the reliability of LRA estimation. When estimating the LRA for a neuron, its true response to one stimulus is held out as test data. Using the LRA fitted on the remaining faces, we can compare the model-predicted MFR with the actual MFR on the held-out data point. For each neuron, the same process is repeated for every face image (as the test data point). We then the correlation coefficient between the true MFR and model-predicted MFR across all held-out data. The neuron is retained for further analysis if the correlation is significant ($p < 0.05$).

Subspace Comparison

For two vector spaces A and B, let $\{a_1, a_2, \ldots a_n\}$ be a set of vectors in A. The explanatory strength of space B for vector a_1 is determined by the percentage of variance of data from space A explained by the best linear combination of Bs basis vectors:

$$R^2 = 1 - \frac{\sum_i (z_i - z_{approx})^2}{\sum_i (z_i - \bar{z})^2},$$

where z_i is the data projection on vector a_1, $\bar{z}$ is the mean, and z_{approx} is the projection to space B.

Discussion

Our results indicate that, while macaque face patch neurons are primarily tuned to combinations of facial features that are rather different from those most important for human social trait perception, one can easily go back and forth using a simple linear operation (linear decoding scheme). There is no particular reason to expect that monkey face patch neurons, or monkeys themselves should particularly care about social trait perception of human faces. However, our results suggest that human social perception of faces may arise simply as linear decoding of featural information in a neural representational system that humans and monkeys share with each other, and with our common primate ancestors.

Leveraging computational modeling, our work represents a novel way to *retroactively* analyze social perceptual information or other face-related cognitive or perceptual information in monkey neural recording data, even if no social ratings are collected for the face images that the monkeys actually saw. We can also easily extend this framework to other kinds of animal neural data, or to human neural recording (or neuroimaging) data, obtained while experimental participants viewed face images. Technologically, this approach presents a promising approach for extracting much more information out of neural data about the neural basis of face processing, than has been hitherto possible.

Acknowledgments

We thank Doris Tsao and Winrich Freiwald for sharing the monkey neural data, Samer Sabri for helpful advice with the writing, and the UCSD CRES program for partial funding.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013, November). The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.*, *142*(4), 1323–1334.

Chang, L., & Tsao, D. Y. (2017, June). The code for facial identity in the primate brain. *Cell*, *169*(6), 1013–1028.e14.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*(6), 681–685.

Freiwald, W. A., & Tsao, D. Y. (2010, November). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Guan, J., Ryali, C., & Yu, A. J. (2018, July). *Computational modeling of social face perception in humans: Leveraging the active appearance model.*

Olivola, C. Y., Funk, F., & Todorov, A. (2014, November). Social attributions from faces bias human choices. *Trends Cogn. Sci.*, *18*(11), 566–570.

Valentine, T. (1991, May). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q. J. Exp. Psychol. A*, *43*(2), 161–204.

Willis, J., & Todorov, A. (2006, July). First impressions: making up your mind after a 100-ms exposure to a face. *Psychol. Sci.*, *17*(7), 592–598.