A computational analysis of gender differences in face-based perception of trustworthiness and dominance

Christine H. Lind (clind@ucsd.edu)
Department of Electrical & Computer Engineering, University of California San Diego
La Jolla, CA 92093 USA

Angela J. Yu (angela@angelayu.org)
Centre for Cognitive Science & Hessian AI Centre, Technical University of Darmstadt
64285 Darmstadt, Germany
Halicioglu Data Science Institute, University of California San Diego
La Jolla, CA 92093 USA

Abstract

Previous work indicates perceived dominance and trustworthiness are both positive predictors for a candidate's electoral and employment success. Interestingly, compared to male faces, female faces exhibit a much stronger anti-correlation between perceived trustworthiness and dominance. Together, these two phenomena place women at a distinctive disadvantage in electoral and general work settings. In this study, we conduct a computational analyses to examine the provenance of the anti-correlation between perceived dominance and trustworthiness in female faces. By identifying and quantifying the facial features that contribute to each social trait in each gender, we find that the perceived female anti-correlation stems predominantly from components unique to female faces. For example, the lip region is a major contributor: visualization of the critical facial features shows that the corners of the mouth move up and down in opposite, i.e. anti-correlated, directions for perceived trustworthiness and dominance for female faces, contributing to the anti-correlation; but they curve in orthogonal, i.e. uncorrelated, directions for male faces. Furthermore, we find that female dominance and trustworthiness perception are correlated in opposite directions along most demographic dimensions, such as gender, age, and race-related dimensions. In contrast, perceived male dominance and trustworthiness are uncorrelated along gender- and age-related dimensions, and postively correlated in the same direction along race-related dimensions. Overall, the results indicate that perceived sexual dimorphism strongly drives the anti-correlated perception of dominance and trustworthiness in female faces (more feminine faces are viewed as more trustworthy and less dominant), while the inconsequence of sexual dimorphism in male trustworthiness perception (despite more masculine faces appearing more dominant) means little anti-correlation of trustworthiness and dominance perception in male faces.

Keywords: face perception; facial features; gender; dominance; trustworthiness; social decision making; social trait ratings; sexual dimorphism

Introduction

Gender bias continues to pose a challenge to women's career advancement in politics and other workplace settings, with opinions appearing to reflect subjective biases rather than factual differences. For instance, in the 2016 U.S. presidential election, Hillary Clinton was criticized for being "too ambitious", "overly dominant", and more "untrustworthy" and "deceitful" than her male opponent, Donald Trump, despite a lack of factual evidence supporting such claims (Foran, 2016). Several psychological studies have also found that only female candidates are penalized when perceived as dominant or power-seeking (Williams & Tiedens, 2016; Okimoto & Brescoll, 2010), a social dilemma known as the "backlash effect" (Rudman, Moss-Racusin, Phelan, & Nauts, 2012) for women who violate prescriptive stereotypes of femininity.

While a number of studies have examined social impressions and perceptual judgments as important factors underlying gender bias in politics and work settings, few have computationally examined the explicit role of face perception (Oh, Buck, & Todorov, 2019), a natural process humans use to form impressions about strangers (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Indeed, humans have been found to form rapid social impressions about character, including traits such as trustworthiness, dominance, attractiveness, and competence, based on face appearance within as little as 100 milliseconds of exposure (Willis & Todorov, 2006). Although the veracity of such face-based social trait impressions is highly debated (Todorov et al., 2015), there is no doubt these perceptions do exist and play an role in consequential social decisions (Todorov et al., 2015). For example, in politics and business, perceived competence, dominance, sociability, and trustworthiness have been shown to be reliable predictors for an individual's electoral and employment success (Todorov et al., 2015; Olivola, Funk, & Todorov, 2014; Olivola & Todorov, 2010). Similarly, trustworthy-looking faces have also been found to be more likely to attract investments and less likely to be convicted in criminal trials (Olivola et al., 2014).

Previously, it was found that trustworthiness and dominance perception of female faces are highly anti-correlated, while the same is only mildly anti-correlated for male faces (He & Yu, 2021). Moreover, it was also found that dominance and trustworthiness perception both positively contribute to the impression of electoral and job candidates for both genders (He & Yu, 2021). Together, these results suggest that the stronger anti-correlation place female candidates, compared to males, at a distinct disadvantage in such work settings (He & Yu, 2021). Moreover, it was found this female disadvantage is due to human observers applying different criteria to the social judgment of female faces compared to that of male faces (He & Yu, 2021). In this work, we conduct a more in-depth analysis of what these gender-specific criteria are, in particular how they depend on the underlying facial features. In the following, we first examine how different facial features contribute to trustworthiness and dominance judgement of both male and female faces, as well as why these contributions lead to a stronger anti-correlation for female faces.

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Then, we also conduct a correlational analysis of social trait and demographic trait perception, combining it with facial feature analysis to arrive at an integrated interpretation.

Results

To computationally investigate which parts of the face contribute to social trait judgment and how, we need an interpretable feature vector representation of faces. To do so, we use a computer vision algorithm known as the Active Appearance Model (Cootes, Edwards, & Taylor, 2001), previously shown to successfully model human (Guan, Ryali, & Yu, 2018; Ryali, Goffin, Winkielmann, Yu, 2021) and monkey (Change & Tsao, 2017) perception of faces. We then conduct a regression analysis of how facial features drive human judgments for each social trait. Previous work shows that linear regression of AAM features can well capture social trait judgments of humans faces (Guan, et al, 2018; Ryali et al, 2021; He & Yu, 2021), and moreover does a fine job replicating the stronger anti-correlation between trustworthiness and dominance perception in female faces (-0.42, $p < 0.001$ in both human ratings and regression model predictions) than in male faces (-0.162, $p < 0.001$ in both).

To quantify how much of the overall anti-correlation (between T and D) for each of female and male faces depend on their unique and shared facial featural components, we decompose the (anti-)correlation between female dominance (D_F) and trustworthiness (T_F) into four components: (1) due to the correlation in facial features unique to female faces, (2) due to the correlation between uniquely female dominance perception and male trustworthiness perception, (3) due to the correlation between uniquely female trustworthiness perception and male dominance perception, (4) due to the correlation between male dominance and trustworthiness perception, to the extent it is also relevant for female faces. The correlation coefficients for each of these components, as well as the percentage quantification of how much each contributes to the overall female trustworthiness and dominance perception, are shown in the first row of Figure 1, across the four columns, respectively. The analogous analysis of male trustworthiness and dominance perception, decomposed into these four components, can be seen in the same row.

Firstly, we find that each of these components has a negative correlation coefficient (c.c.), indicating a general tendency for trustworthiness and dominance perception in both genders with respect of each of the components. Secondly, we find that components of the anti-correlation unique to each gender (first column) account for the majority of anti-correlation for both genders (83% for female, 71% for male). Thirdly, we find that female trustworthiness and dominance perception "inherits" very little from male trustworthiness and dominance perception (4%), while male trustworthiness and dominance perception "inherits" a substantial amount in the converse direction (24%). Altogether, this suggests that the much stronger anti-correlation in female trustworthiness and dominance perception is unique to female faces.

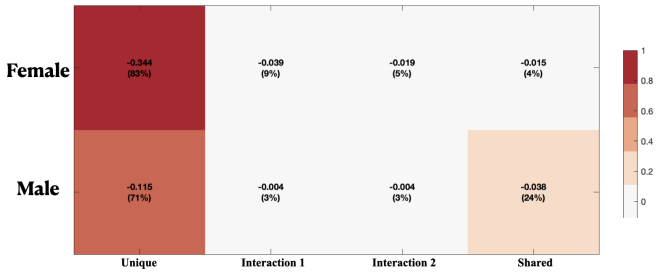

Figure 1: Unique (left column) and shared (right column) components, as well as the interactions between the two components (middle columns) of the dot product for both female (top row) and male (bottom row) faces. Percentage is the correlation of the component divided by the total correlation for each gender. See Methods for how these components were determined.

So far, we have decomposed the correlational analysis into statistical components unique and shared between the genders; next, we use a different a method to characterize and better visualize how the overall perceptual correlation decompose into different facial regions. As illustrated in Figure 2A, we decompose the face into different face regions delineated by the landmarks of AAM – eyebrows, eyes, nose, lips, upper face (excluding eyebrows, eyes, and nose), lower face (excluding lips). Using the fitted regression models, we can use only the AAM features corresponding to each facial region to predict social traits (trustworthiness and dominance), and compute the dot product between the two traits for different facial regions (see Figure 2B), separately for female (top row) and male (bottom row) faces. Note that, as the data is standardized, the overall correlation coefficient is equal to the dot product between the two traits. We thus decompose the correlation coefficient in terms of the dot product for different face regions to investigate how different facial features contribute to the overall anti-correlation for male and female faces. We find that for female faces, while every face region contributes to the overall anti-correlation between trustworthiness and dominance ratings (the model-predicted correlation is always negative for each region), the two most prominent facial regions are lips ($\rho = -0.096$, 23% of overall anti-correlation) and lower face excluding lips ($\rho = -0.084$, 20%). In contrast, the weaker anti-correlation between trustworthiness and dominance predicted by the regression models for male faces are most prominently driven by facial features in the upper face ($\rho = -0.074$, 46%), eyebrows ($\rho = -0.037$, 23%), and nose ($\rho = -0.036$, 22%). In contrast to female faces, the lips region of the model predicts very little anti-correlation in male faces ($\rho = -0.016$, 10%), and the lower face excluding lips actually predicts a positive correlation ($\rho = 0.014$, -9%), thus canceling out negative correlation in other parts of the face.

Similarly, we can use the facial region decomposition method to quantify whether and how much each region con-

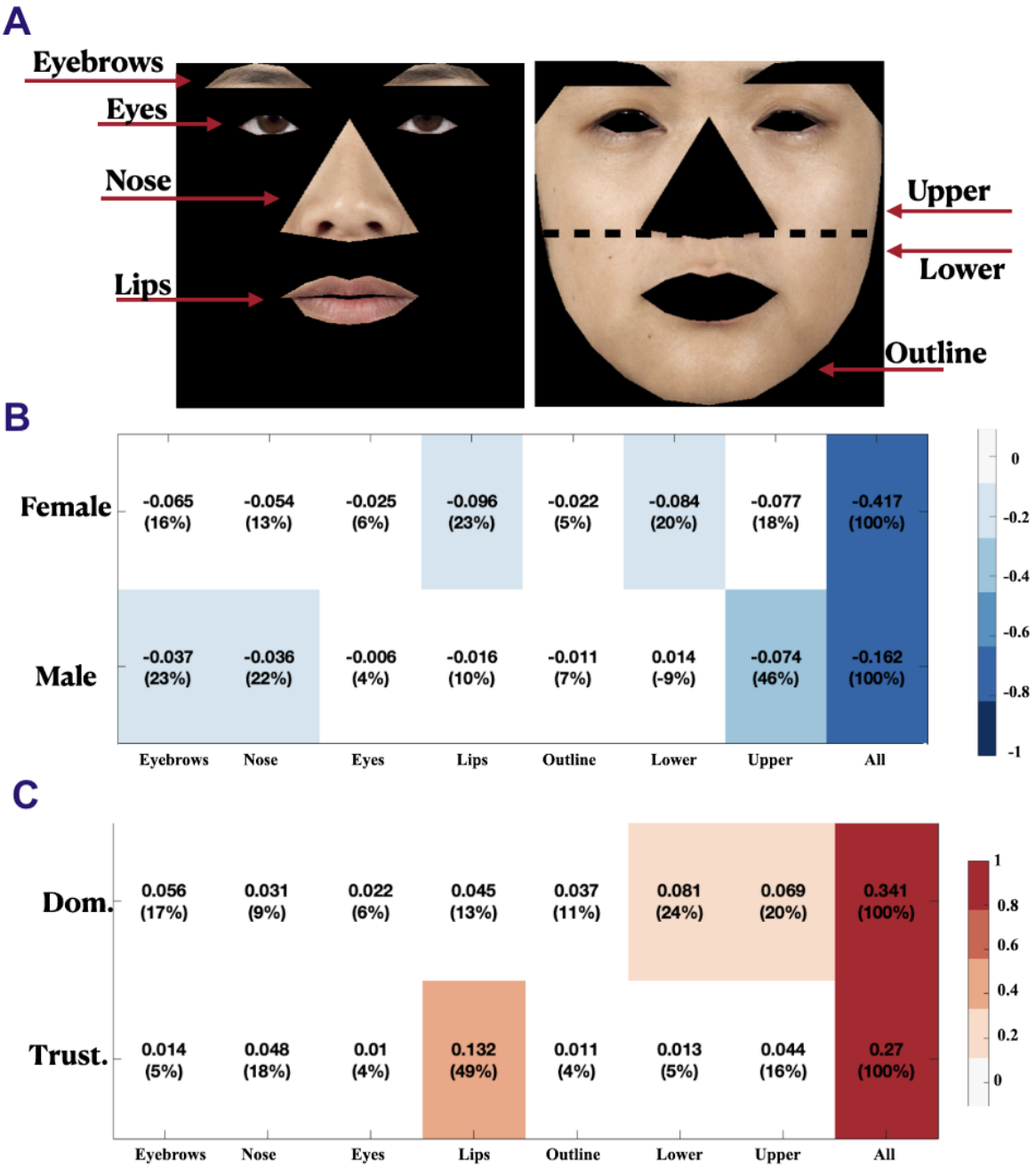

Figure 2: Decomposition of facial feature correlation into facial regions. **A**: Illustration of the seven facial regions: eyes, eyebrows, nose, lips, upper, lower and face outline. **B**: Components of the dot product between dominance and trustworthiness decomposed into facial regions for female (top row) and male (bottom row) faces. **C**: Components of the dot product decomposed into facial regions for female and male dominance (top row) as well as female and male trustworthiness (bottom row). For both B and C, the overall dot product (last column) is the total correlation, and percentage is the dot product of the facial region divided by total correlation.

tributes to the overall correlation between male and female faces in the perception of dominance (Figure 2C top row) and trustworthiness (Figure 2C bottom row). Firstly, we see that that the perception of dominance and trustworthiness are in fact significantly though not very highly correlated between male and female faces ($\rho = 0.341$, $\rho = 0.270$, respectively, $p < 001$). Moreover, every facial region contributes positively to this positive correlation, indicating that whatever featural dependence underlying the differential anti-correlation between the two genders are not some extremely overt differences, such as negative region-specific correlations, but rather subtle differences that are difficult to see based on a pure correlational analysis. In particular, the lips region is positive between males and females for both trustworthiness ($\rho = -0.132$), and dominance ($\rho = 0.045$), and yet this region strongly contributes to female anti-correlation while barely contributes to male anti-correlation.

So how exactly do the different facial regions contribute to female and male trustworthiness and dominance perception, and why is there a stronger anti-correlation for female faces? Here, we take advantage of AAM's ability to readily visualize the facial featural changes that most strongly drive human social trait perception (via the linear regression models of trait perception, see Methods and Guan et al, 2018; Ryali et al, 2021). As shown in Figure 3, which shows how texture features change (via pixel value changes in each row) and shape feature changes (indicated by red arrows), the overall judgment of trustworthiness and dominance are similar between male and female faces, consistent with the overall positive correlation between F_T and M_T, and between F_D and M_D. However, there are also clear differences in the details that drive a stronger anti-correlation between the two traits for female faces. For example, consistent with the correlational analysis (Figure 1 and 2) indicating the lip/lower face region playing a particularly important role, one can see that the corners of the mouth curve up and down in opposite directions for F_D and F_T for female faces (resulting in a perceptual anti-correlation), as indicated by red arrows in Figure 3A, but in orthogonal directions for M_D and M_T for male faces (resulting in perceptual non-correlation). Moreover, redder lips (Figure 3B) increase female trustworthiness but decreases dominance, while lip redness appears to play little role in male social trait perception.

Visual inspection suggests that there might be a holistic demographic effect that is difficult to pin down by analyzing individual face regions or features. While dominance in both genders (F_D and M_D) appear negatively correlated with appearing feminine and Asian, while positively correlated with age, F_T appears to have exactly the opposite relationship with appearing feminine, Asian, and older; in contrast, M_T does not seem to have obvious relationship with these demographic trait impressions. To quantify this, we compute the correlation coefficients between social trait ratings and demographic trait ratings (Figure 4). We find that indeed, F_D and F_T are driven in opposite directions along demographic dimensions relating to sexual dimorphism (perceived femininity, perceived masculinity, real gender), (real) age, and race (Asian/non-Asian). In contrast, M_D and M_T are driven in the same direction along race-related dimensions (White/non-White, Black/non-Black), while only M_D is driven by age and sexual dimorphism. This sheds further light on why there is greater anti-correlation in perceived trustworthiness and dominance for female faces: while dominance perception is strongly related to demographic perception in both genders, trustworthiness perception only strongly relate to demographic perception in female faces. In particular, perceived femininity (sexual dimorphism) is strongly weighed in the assessment of female trustworthiness, outweighing all other demographic considerations, while perceived masculinity appears unimportant for male trustworthiness perception. This asymmetry may explain why previous studies have found that counter-stereotyping disadvantages women more than men (Cuddy, Fiske, Glick, 2008).

Discussion

Previous work (He & Yu, 2021) found that gender-specific criteria are used by humans in face-based judgment of social traits, resulting in a far larger anti-correlation between dominance and trustworthiness for female faces, compared to male faces. As a result, women can be expected to be disproportionately negatively affected in situations that require candidates to be perceived as both dominant and trustworthy, including political elections and the workplace (He & Yu, 2021). In this work, we examined in-depth what these gender-based criteria are. We found 1) the stronger female anti-correlation primarily depends on the lips region and the bottom half of the face in general, while the weaker male anti-correlation primarily depends on the upper half of the face, especially the eye brows and nose; 2) differences between female and male correlation depend on rather subtle differences in featural dependence on each region, as there is positive correlation between male and female faces in every facial region for the perception of both trustworthiness and dominance; 3) these subtler featural difference may be related to the much stronger demographic influence on female trustworthiness perception, in particular sexual dimorphism, than on male trustworthiness perception, as the same demographic factors influence dominance perception in the opposite direction to female trustworthiness for both female and male faces.

Our results have an interesting relationship with previous work showing that counter-stereotyping individuals are less positively perceived than gender-stereotype-conforming individuals. For example, dominant-looking female faces are less preferred than both less dominant-looking female faces and dominant-looking male faces (Cuddy et al, 2008). This has been interpreted as evidence that there is a "backlash effect" toward counter-stereotyping individual. However, this bias effect curiously has been found to be much stronger for women than men. For example, trustworthy-looking male faces are evaluated more positively than trustworthy-looking

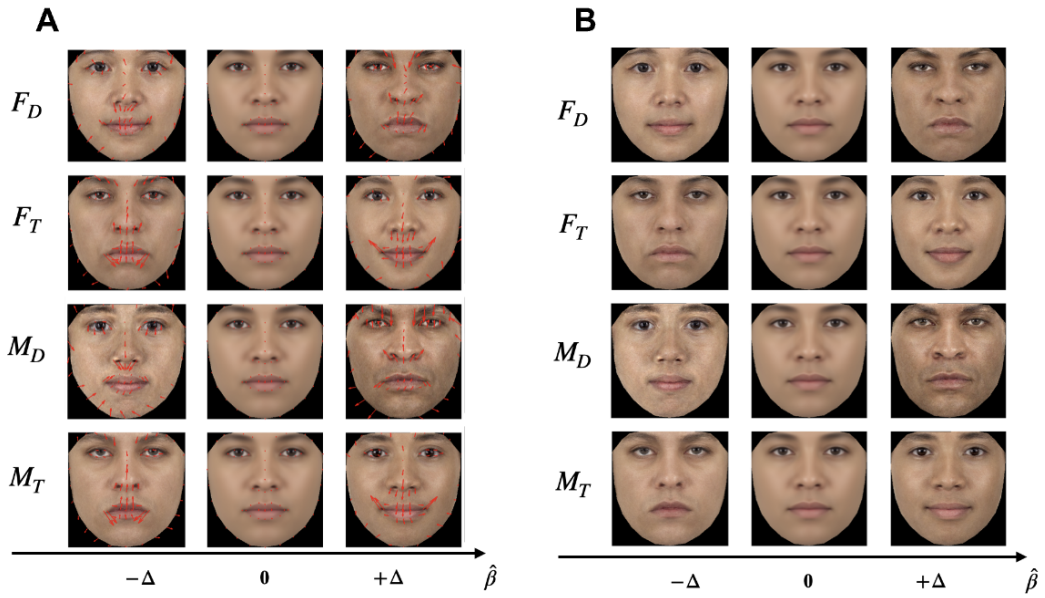

Figure 3: **A**: Neutral faces (middle column) visualized to look less/more (left/right column) dominant and untrustworthy (dominant: rows 1,3; untrustworthy: rows 2,4) according to the female and male models (female: rows 1-2; male: rows: 3-4). Arrows indicate the movement of shape features. **B**: Same as **A** but without shape features for easier visualization of texture features.

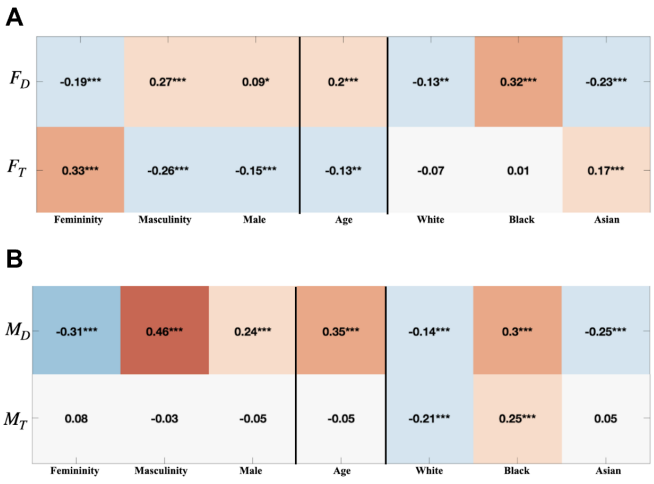

Figure 4: Correlation between perceived dominance and trustworthiness and demographic information for female (**A**) and male (**B**) faces. *: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001.

female faces (Cuddy et al, 2008), even though trustworthiness is thought to be gender-conforming for women, while the dominance is gender-conforming for men (Sutherland, Young, Mootz, Oldmeadow, 2015). Our work suggests that the answer to this apparent mystery is that female trustworthiness perception depends on sexual dimorphism, while male trustworthiness perception does not; on the other hand, dominance perception in both genders positively correlate with masculinity and negatively with femininity perception. Thus, dominant-looking females are consequently viewed as less trustworthy, while trustworthy-looking males are not necessarily viewed as less dominant. This asymmetry puts counter-stereotyping women at a greater disadvantage than men.

Our findings about the role of gender information on subsequent face perception is also consistent with the neuroscience literature on the temporal dynamics of face processing, whereby gender information has been found to be available earlier than other types of information (Dobs, Isik, Pantazis, Knwisher, 2019). One limitation of our study is that we divided a face into regions based on landmarks, which resulted in large general areas such as upper and lower halves of the face. We see (Figure 2) that these regions account for a large portion of the anti-correlation between male and female dominance (44% combined). The use of a more sophisticated algorithm, such as a clustering algorithm, might be able to group more naturally correlated regions together, such as chin and cheeks, resulting in more informative regions. Our current method of looking at the contribution of different facial regions also does not account for correlation between these regions. For instance, we do not know if the contribution of

the eye region is independent of the mouth region. As such, we do not know how the interaction of facial features contribute to perceived dominance and trustworthiness. These are fruitful areas of future investigations.

Methods

Dataset

We use the neutral-expression face images in the publicly available Chicago Face Database (CFD) (Ma et al., 2015). This dataset consists of 109 East Asian (57 female), 197 black (104 female), 108 Hispanic (56 female) and 181 white (90 female) faces, along with collected social and demographic trait ratings. All trait ratings and demographic information used in this work are part of the CFD dataset.

Model Descriptions

Active Appearance Model (AAM): AAM is a well-established computer vision model (Cootes et al., 2001) with neural relevance (Chang & Tsao, 2017), consisting of shape and texture features. The shape features are the coordinates of a set of predefined landmarks, while the texture features are the pixel values of each face image after having been warped to the average landmark location. We train the AAM on the CFD faces, then perform additional PCA on the combined shape and texture features. We retain all PCs to obtain a 596 dimensional "face space".

Linear Trait Axis (LTA): The LTA (Guan et al., 2018) for each social trait is computed as the normalized regression coefficients of ratings regressed against standardized AAM features: $y = \beta x$, where y is the standardized ratings for the trait, x is the standardized AAM features for a face, and β is the vector of regression coefficients. The LTA is then defined as the normalized regression vector. The LTA specifies a direction in the face space that (linearly) maximally alters the response of the task. Note that, since we retain all features and the face data is standardized, the correlation between two traits is just the dot product of their LTAs.

Orthogonalization: To extract the components unique to the female model ($\beta_{F\perp M}$), we orthogonalize the female LTA (β_F) against the male (β_M) using a standard orthogonalization procedure:

$$\beta_{F\perp M} = \beta_F - (\beta_M \cdot \beta_F)\beta_M \tag{1}$$

Note that $\beta_M \cdot \beta_F$ is just a constant c, where $\cdot$ denotes the dot product. We can then write the female LTA in terms of its unique component ($\beta_{F\perp M}$) and component shared with male faces (β_M) as:

$$\beta_F = \beta_{F\perp M} + c_1\beta_M \tag{2}$$

Correlation Coefficients: Since we retain all features and the face data is standardized, the correlation between two traits is just the dot product of their LTAs. The correlation between female dominance (D_F)and trustworthiness (T_F) in terms of unique and shared components thus becomes:

$$D_F \cdot T_F = D_{F\perp M} \cdot T_{F\perp M} + c_D D_{F\perp M} \cdot T_M \tag{3}$$

$$+c_T T_{F\perp M} \cdot D_M + c_D c_T D_M \cdot T_M \tag{4}$$

where

$$c_D = F_D \cdot M_D, \quad c_T = F_T \cdot M_T \tag{5}$$

Note that the first term in Equation corresponds to the contribution unique to female faces, the last term corresponds to the contribution derived from male faces, while the middle two terms are interaction terms.

Facial Regions: Each component in the LTAs and regression vectors correspond to a shape or texture feature. We group the components corresponding to different facial regions together to obtain a set of seven facial regions: eyebrows, nose, eyes, lips, face outline, upper half of the face, and lower half of the face. As the data is standardized, the total correlation between two traits can thus be written as a sum of the dot product of these regions. For instance, the total correlation between female dominance (D_F) and trustworthiness (D_T) thus becomes:

$$c.c = \sum_{r\in region} D_F(r) \cdot T_F(r) \tag{6}$$

The percentage contribution of a region i to the overall correlation thus becomes:

$$\frac{D_F(i) \cdot T_F(i)}{\sum_{r\in region} D_F(r) \cdot T_F(r)} \tag{7}$$

which is the dot product between the LTAs of facial region i divided by the overall correlation.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of experimental psychology*, *142*(4), 1323-34.

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

Cootes, T., Edwards, G., & Taylor, C. (2001). Active appearance models. ieee trans. *Pattern Anal*, *23*, 681-685.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, *40*, 61–149.

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Na- ture Communications*, *10*(1), 1-10.

Foran, C. . (n.d.). *Sept 17)*. The curse of hillary clinton's ambition. The Atlantic. Retrieved from https://www.theatlantic.com/

Guan, J., Ryali, C., & Yu, A. J. (2018). Computational model- ing of social face perception in humans: Leveraging the active appearance model. biorxiv(2). , *360776*. Retrieved from https://www.biorxiv.org/content/10.1101/360776v1 doi: 10.1101/360776

He, Z. W., & Yu, A. J. (2021). Gender differences in face-based trait perception and social decision making. In *Proceedings of the annual meeting of the cognitive science society.* 43.

Huang, S. J., Ryali, C. K., Liu, J., Guo, D., Guan, J., Li, Y., & Yu, A. J. (2018). A Model-Based Investigation of the Biological Origin of Human Social Perception of Faces. Retrieved from `https://www.faceplusplus.com`

Ma, D., Correll, J., & Wittenbrink, B. . (n.d.). 01). *The chicago face database: A free stimulus set of faces and norming data*, *47*.

Meco, L. D. . (n.d.). *Dec 17).* Missing from the Conversations about Tech and Elections? Women. Retrieved from `https://www.genderontheballot.org/`

Oh, D. W., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, *30*(1), 65-79.

Okimoto, T. G., & Brescoll, V. L. (2010). The price of power: Power seeking and backlash against female politicians. *Personality and Social Psychology Bulletin*, *36*(7), 923-936.

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

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and vot- ing. *Journal of Nonverbal Behavior*, *34*(2), 83-110.

Paul, D., & Smith, J. L. (2008). Subtle sexism? examining vote preferences when women run against men for the presidency. *Journal of Women, Politics and Policy*, *29*(4), 451-476.

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prej- udice against female leaders. *Journal of Experimental Social Psychology*, *48*(1), 165-179.

Ryali, C., Wang, X., & Yu, A. J. (2020). Leveraging computer vision face representation to understand human face representation. In *Proceedings of the 42th cognitive science society conference.*

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

Sutherland, C. A., Young, A. W., & Mootz, C. A. a. (2015). old- meadow, j. In *A* (p. 186-208). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical fe- male faces are negatively evaluated. British Journal of Psychology, 106(2.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determi- nants, consequences, accuracy, and functional signifi- cance. *Annual Review of Psychology*, *66*, 519-545.

Williams, M. J., & Tiedens, L. Z. (2016). The subtle suspension of backlash: A meta-analysis of penalties for women's implicit and explicit dominance behavior. *Psychological Bulletin*, *142*(2), 165.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, *17*(7), 592-598.