Supplemental Material

Gender biases in impressions from faces:

Empirical studies and computational models

DongWon Oh[1]*, Ron Dotsch[2], Jenny Porter[1], Alexander Todorov[1]

[1]Department of Psychology, Princeton University, Princeton, NJ 08544

[2]Department of Psychology, Utrecht University, Utrecht, The Netherlands

*Correspondence: DongWon Oh (dong.w.oh@gmail.com)

Table of Content

Underlying Structure of Male and Female Face Impressions in Study 1

Principal Components Analysis Procedure

When summarizing and visualizing the results of the orthogonal PCAs in Studies 1a and 1b, we followed the Kaiser rule: For both male and female faces, we reported the first two components, i.e., PC1 and PC2, because they had eigenvalues bigger than 1 (Table S1). This indicates that the third and following components were unable to explain a variance larger than a single input variable (i.e., a trait rating) alone, so we deemed it as unnecessary to include them. For consistency, in Study 1c we restricted the number of components to two although the PCA solution found four components with eigenvalue > 1 for both face genders (see Table S1 for details). However, it should be noted that the 14 trait impressions used in Studies 1a and 1b were selected in a data-driven fashion after recoding of free-response descriptions of person impressions from faces (Oosterhof & Todorov, 2008), whereas there is no report that the 15 traits used in Study 1c were chosen in a data-driven way (Ma, Correll, & Wittenbrink, 2015).

Comparison between 2008 and 2018 Face Evaluation Structures

In addition to the main analyses, we examined the similarity between the dataset of Oosterhof & Todorov (2008; Study 1a) and the newly collected data (Study 1b). When face genders were collapsed across, the two datasets led to highly similar PCA solutions, suggesting the same structure of impressions between the two datasets collected about ten years apart (see the main text for details). When face genders were considered and separate PCAs were conducted for male and female ratings, the two datasets led to highly similar PCA solutions again, suggesting the gender specific structures of impressions from the dataset of Oosterhof and Todorov (2008) and from the present dataset (Study 1b) are highly similar. Specifically, we ran correlational analyses of the PCA loadings of the trait ratings on PC1 and PC2 across the two datasets. A high correlation between the PCA loadings of the two datasets indicates high similarity between the impression structures. Between the 2008 data and the current data, the component loadings of the traits were highly similar for both male (R = .97) and female face impressions (R = .98), suggesting high stability of the impressions of male and female faces over time.

Analysis of the Effects of Raters' Gender Stereotype Endorsement

When testing for the effect of the raters' GSE on facial impressions (see Table S2 for the traits used in the GSE questionnaire), we conducted an additional analysis on the structure of impressions using four factor scores of GSE (each of which represents specific subtypes of gender stereotypes) in addition to the analyses reported in the main text (Study 1b). In the analyses reported in the main text, we used the GSE score, the sum of each rater's responses to all items in the questionnaire (see the main text for details). The additional analysis yielded consistent results with the main analyses.

To examine if the relationship between the impression differentiation and the rater GSE is affected by gender- or valence-specificity of stereotypes, we computed four factor scores for each rater – stereotype gender [male/female] × stereotype valence [positive/negative]. Each factor score represented the extent to which each rater supported gender- and valence-specific stereotypes. The four-factor confirmatory factor model revealed an acceptable albeit minimum level of fitness ($\chi^2(164) = 826.21$, $P < .001$; CFI = 0.827; RMSEA = 0.093, 90% CI = [0.087,0.099]; SRMR = 0.067). The resulting factor loadings of this four-factor solution showed the expected relationships between the factors and the questionnaire responses (see Table S2 for details).

We then simply replicated the analyses reported in the main text using each GSE factor score rather than the GSE sum score. Although we explored the relationship between impression differentiation and the rater GSE in relation to the specific subtypes of stereotype (gender × valence), people's attitudes towards the four subtypes (i.e., male and positive, male and negative, female and positive, and female and negative) go hand in hand (Glick & Fiske, 1996; Glick et al., 2000, 2004). That is, if one holds one subtype of gender stereotypes, say, stereotypes about stereotypically male and positive traits (e.g., "men are more analytical than women"), then the person is likely to hold the other three gender stereotype subtypes as well (e.g., "men are more hostile than women", "women are more gullible than men", "women are more nurturing than men"). Thus, we did not expect any uniquely distinct effect of the stereotype subtypes on the effect of GSE on impression differentiation. As we expected, the results for all four subtypes (factor scores) were consistent with the results reported in the main text. First, when a rater GSE factor score increased, regardless of the

subtype of GSE, the correlation between impressions became stronger for both male and female faces, and the quadratic regression model explained more variance than the linear regression model (Fig. S2). Second, female face impressions had stronger inter-correlations and larger amount of variance explained by valence than male face ratings did (see below and Figure S2 for details).

Role of Raters' Gender Stereotypes about Male × *Positive Traits*. The linear regression model was significant for the ratings of both face genders (male faces: R^2 = .93, F(1,136) = 1713.94, P < .001; female faces: R^2 = .79, F(1,136) = 515.78, P < .001), but the quadratic model (male faces: R^2 = .94, F(2,135) = 1070.56, P < .001; female faces: R^2 = .86, F(2,135) = 428.11, P < .001) explained significantly more variance (male faces: F(1,136) = 32.33, P < .001; female faces: F(1,136) = 71.83, P < .001). Correspondingly, the amount of variance in the ratings explained by the valence component (PC1) followed the same quadratic pattern of change across the male × positive GSE factor score: The linear regression model was significant (male faces: R^2 = .95, F(1,136) = 2468.34, P < .001; female faces: R^2 = .84, F(1,136) = 717.14, P < .001), but the quadratic model (male faces: R^2 = .95, F(2,135) = 1343.58, P < .001; female faces: R^2 = .88, F(2,135) = 496.27, P < .001) explained a larger amount of variance than the linear models did (male faces: F(1,136) = 12.38, P < .001; female faces: F(1,136) = 44.74, P < .001). Across the factor score, female face ratings had higher correlational coefficients (ts > 6.32, Ps < .001) and larger amount of variance explained by the valence component than male face ratings (ts > 12.87, Ps < .001).

Role of Raters' Gender Stereotypes about Male × *Negative Traits*. The linear regression model was significant for the ratings of both face genders (male faces: R^2 = .84, F(1,136) = 712.15, P < .001; female faces: R^2 = .89, F(1,136) = 1076.63, P < .001), but the quadratic model (male faces: R^2 = .91, F(2,135) = 697.08, P < .001; female faces: R^2 = .96, F(2,135) = 1442.86, P < .001) explained significantly more variance (male faces: F(1,136) = 110.20, P < .001; female faces: F(1,136) = 203.78, P < .001). Correspondingly, the amount of variance in the ratings explained by valence followed the same quadratic pattern of change across the male × negative GSE factor score: The linear regression model was significant (male faces: R^2 = .83, F(1,136) = 648.12, P < .001; female faces: R^2 = .93, F(1,136) = 1725.88, P < .001), but the quadratic model (male faces: R^2 = .90, F(2,135) = 579.89, P < .001; female faces: R^2 = .97, F(2,135) = 2129.94, P < .001) explained a larger amount of

variance than the linear models did (male faces: $F(1,136) = 89.57$, $P < .001$; female faces: $F(1,136) = 186.02$, $P < .001$). Across the factor score, female face ratings had higher correlational coefficients ($ts > 7.43$, $Ps < .001$) and larger amount of variance explained by valence than male face ratings ($ts > 14.35$, $Ps < .001$).

Role of Raters' Gender Stereotypes about Female × *Positive Traits*. The linear regression model was significant for the ratings of both face genders (male faces: $R^2 = .91$, $F(1,136) = 1345.46$, $P < .001$; female faces: $R^2 = .75$, $F(1,136) = 399.19$, $P < .001$), but the quadratic model (male faces: $R^2 = .95$, $F(2,135) = 1387.28$, $P < .001$; female faces: $R^2 = .87$, $F(2,135) = 449.27$, $P < .001$) explained significantly more variance (male faces: $F(1,136) = 132.10$, $P < .001$; female faces: $F(1,136) = 127.64$, $P < .001$). Correspondingly, the amount of variance in the ratings explained by valence followed the same quadratic pattern of change across the female × positive GSE factor score: The linear regression model was significant for both genders (male faces: $R^2 = .88$, $F(1,136) = 953.66$, $P < .001$; female faces: $R^2 = .78$, $F(1,136) = 482.23$, $P < .001$), but the quadratic model (male faces: $R^2 = .95$, $F(2,135) = 1270.04$, $P < .001$; female faces: $R^2 = .89$, $F(2,135) = 535.72$, $P < .001$) explained a larger amount of variance than the linear models did (male faces: $F(1,136) = 198.88$, $P < .001$; female faces: $F(1,136) = 130.40$, $P < .001$). Across the factor score, female face ratings had higher correlational coefficients ($ts > 10.60$, $Ps < .001$) and larger amount of variance explained by valence than male face ratings ($ts > 17.16$, $Ps < .001$).

Role of Raters' Gender Stereotypes about Female × *Negative Traits*. The linear regression model was significant for the ratings of both face genders (male faces: $R^2 = .93$, $F(1,136) = 1715.09$, $P < .001$; female faces: $R^2 = .56$, $F(1,136) = 176.45$, $P < .001$), but the quadratic model (male faces: $R^2 = .93$, $F(2,135) = 954.42$, $P < .001$; female faces: $R^2 = .64$, $F(2,135) = 121.59$, $P < .001$) explained significantly more variance (male faces: $F(1,136) = 15.16$, $P < .001$; female faces: $F(1,136) = 29.61$, $P < .001$). Correspondingly, the amount of variance in the ratings explained by valence followed the same quadratic pattern of change across the female × negative GSE factor score: The linear regression model was significant (male faces: $R^2 = .94$, $F(1,136) = 2115.56$, $P < .001$; female faces: $R^2 = .63$, $F(1,136) = 233.82$, $P < .001$), but the quadratic model (male faces: $R^2 = .95$, $F(2,135) = 1258.17$, $P < .001$; female faces: $R^2 = .70$, $F(2,135) = 157.86$, $P < .001$) explained a larger amount of variance than the linear models did (male faces: $F(1,136) = 25.15$, $P < .001$; female faces: $F(1,136) =$

30.75, $P < .001$). Across the factor score, female face ratings had higher correlational coefficients (ts > 2.14, Ps < .033) and larger amount of variance explained by valence than male face ratings (ts > 7.69, Ps < .001).

It should be noted that in all four cases, although the quadratic models explained significantly more variance than the linear models, the magnitude of the quadratic effects was much smaller than the magnitude of the linear effects, and the increase in the intercorrelations of trait ratings as a function of GSE was largely monotonic (Fig. S2).

Analysis of the Effects of Rater Gender

When testing for the effect of raters' gender on facial impressions, we conducted an additional analysis using a 2 [face gender] × 2 [rater gender] repeated measures ANOVA on the absolute values of the inter-impression correlational coefficients in addition to the analyses reported in the main text (Study 1b). In the analyses reported in the main text, we used Jennrich (1970) tests of matrix equality (see the main text for details) instead of an ANOVA because the dataset violates the assumption of sample independence. However, given that ANOVA is known to be rather robust to violations of independence, we report the additional result below in the *Supplemental Material*. The additional analysis yielded consistent results with the main analyses (we calculated a generalized eta-squared (η_G^2) as the measure of the effect size of each effect (Olejnik & Algina, 2003) to account for the repeated measures design).

The 2 [face gender] × 2 [rater gender] repeated measures ANOVA on the absolute values of the intercorrelational coefficients between trait ratings yielded a significant effect of the rater gender ($F(1,90) = 24.36$, $P < .001$, $\eta^2_G = .03$) with the female raters showing a higher level of correlations between trait ratings ($M_{|r|} = 0.59$, $SD_{|r|} = 0.23$) than the male raters ($M_{|r|} = 0.50$, $SD_{|r|} = 0.23$), indicating that female raters had less differentiated face impressions. This main effect was qualified by a significant interaction between the rater gender and face gender ($F(1,90) = 4.87$, $P < .05$, $\eta^2_G = .01$). Female raters showed a stronger cross-trait intercorrelations for female ($M_{|r|} = 0.61$, $SD_{|r|} = 0.21$) than for male faces ($M_{|r|} = 0.56$, $SD_{|r|} = 0.25$; $t(90) = 2.50$, $P = .01$, Bonferroni correction), whereas male raters showed the same level of cross-trait correlations for female ($M_{|r|} = 0.51$, $SD_{|r|} = 0.21$) and for male faces ($M_{|r|} = 0.50$, $SD_{|r|} = 0.25$; $t(90) = 0.41$). This finding suggests that the less differentiated impressions of female faces are primarily due to female raters.

Building of Male and Female Face Impression Models in Study 2

A data-driven face modeling approach allows one to build models of impressions without a priori assumptions about the effect of specific facial features (e.g., the size of the nose) for impressions (Dotsch & Todorov, 2012; Funk, Walker, & Todorov, 2016; Gosselin & Schyns, 2001; Jack & Schyns, 2017; Mangini & Biederman, 2004; Oosterhof & Todorov, 2008; Todorov, Dotsch, Wigboldus, & Said, 2011; Walker & Vetter, 2009; 2016). In the standard, hypothesis-driven approach, different facial features are manipulated. However, the combinations of features rapidly proliferate as the number of features increases (Jack & Schyns, 2017; Todorov et al., 2011), damaging the feasibility and/or the statistical power of the investigation. For example, a simple factorial design for the investigation of the effect of even only ten binary facial features (e.g., a long vs. short nose) would lead to 2^{10} experimental conditions. The data-driven approach prevents this by presenting a relatively small number of faces (e.g., 300 faces), which randomly vary in their features.

In Studies 2–3, we used the statistical face space model of FaceGen 3.2 (Singular Inversions) that captures the variance from a large sample of real human faces with 100 orthogonal dimensions (Todorov et al., 2011; Todorov & Oosterhof, 2011). Each dimension represents the variance in a holistic combination of features. A single face is represented as a vector in the statistical space (i.e., an array of 100 numbers).

Using this approach, one can generate an unlimited number of faces by randomly sampling parameters and generating the corresponding faces as images. Participants then judge the randomly sampled faces on a trait of interest, e.g., trustworthiness. One can model the trait judgment by extracting the change in face parameters that are correlated with the change in the trait judgment. With the resulting model, we can visualize what aspects of facial appearance change when an impression of the trait changes.

The trait model can be applied to any new face to make it appear more or less trait-like (e.g., trustworthy) by moving its corresponding parameters across the modeled judgment. With these manipulated faces, one can study what types of facial cues (e.g., emotional facial gestures, perceived physical strength) predict the perceived level of the trait. In addition, a data-driven statistical face model allows one to vary a particular perceived trait of faces while specifically controlling for another trait (Oh, Buck, & Todorov, 2019; Todorov, Dotsch, Porter, & Oosterhof, 2013). For example, Oh and colleagues (2019) manipulated the

perceived competence of faces while controlling for facial attractiveness, thereby effectively suppressing the halo effect underlying competence impressions. This procedure found that facial masculinity is one of the ingredients of competence impressions, revealing gender biases in competence impressions.

Validation of Male and Female Face Impression Models in Study 3

To create the face stimuli for the validation studies, we manipulated their level of perceived trustworthiness and dominance. We added -3, -2, -1, 0, 1, 2, and 3*SD*s to the trustworthiness or dominance value of the 25 randomly generated male and 25 randomly generated female faces, using either the male or the female model. In other words, we moved the coordinates of the faces in the face space along one of the four gender-specific trait models.

This was a different approach from that of previous validation studies, which did not control for the gender of the faces: Todorov and colleagues (2013) manipulated the trait dimension of randomly generated faces to take specific values (i.e., -3, -2, -1, 0, 1, 2, and 3*SD*s on each trait model). These procedures are inappropriate when validating gender-specific trait models that are inherently correlated with gender in raters' perception. For instance, male faces are perceived as more dominant than female faces, and female faces are perceived as more trustworthy than male faces (e.g., Sutherland et al., 2013; Studies 1a and 1b in the main text). Because of these correlations, these procedures would decrease gender-related differences between male and female face sets, as they would project all the faces, regardless of their gender, onto the trait dimensions with the same values. As a result, we would essentially be generating less male-like male faces and less female-like female faces for the validation stimulus set.

Note that the parameters of the average male face and the parameters of the average female face used here were based on samples of actual male and female faces. That is, 3D laser scans of these male and female faces were used to construct the FaceGen statistical face space and extract the 100 face parameters. In the current project, by adding dimension values of [-3 to 3*SDs*] to the original faces (rather than assigning the faces to the dimension values as in the previous approach), we maintained the gender-related facial information in the male and female faces.

When cross-validating gender-specific impression models using ANOVAs, we calculated a generalized eta-squared (η_G^2) as the measure of the effect size of each effect (Olejnik & Algina, 2003) to account for the repeated measures design of the experiment.

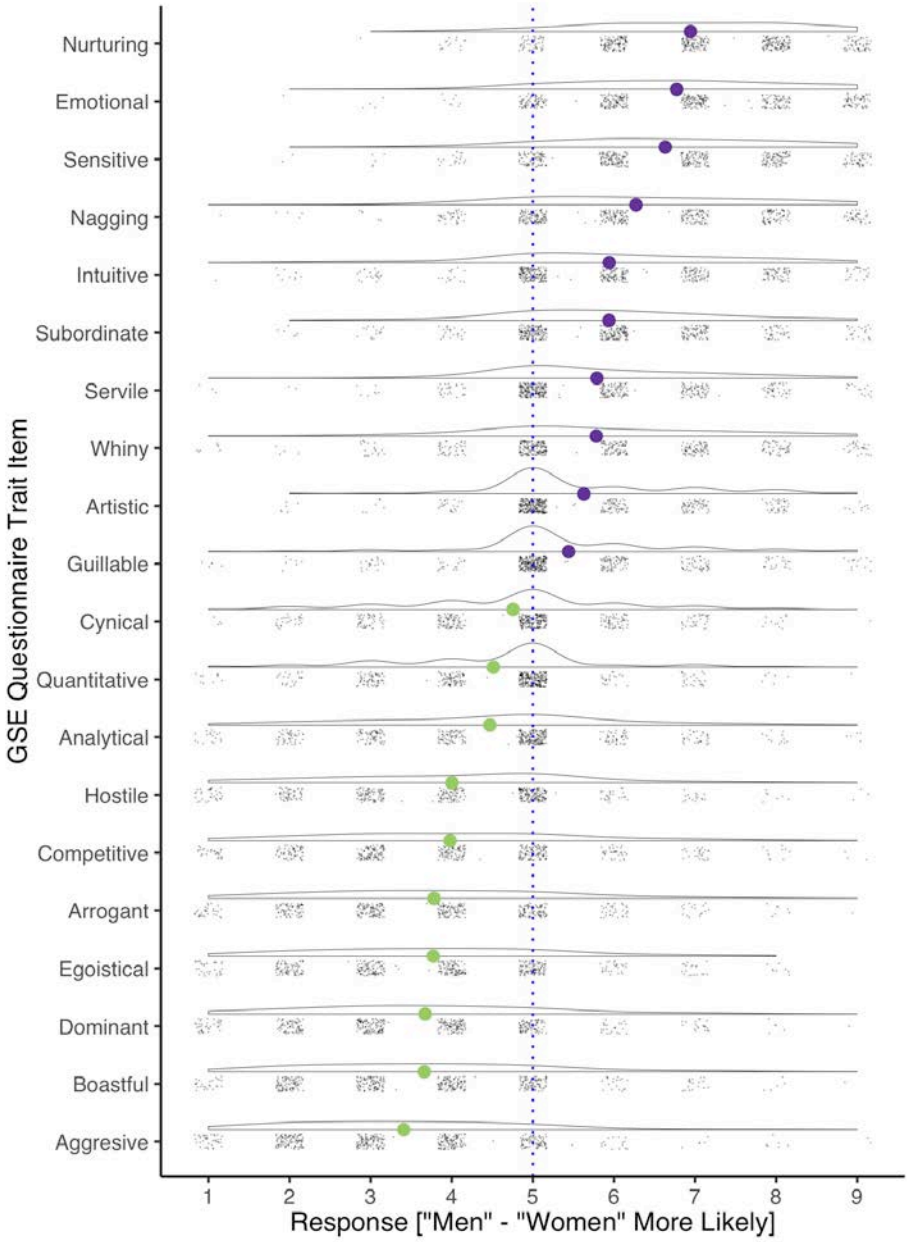

Figure S1. The distribution of raters' responses to gender stereotype endorsement (GSE) questions in Study 1b. In every question, raters showed a bias away from the middle score (blue dotted line) in the direction consistent with gender stereotypes (*t*s > 7.02, *P*s < .001). The bigger purple (traits associate with women) and green dot (traits associate with men) denote the mean response, and the smaller black dots raw responses. All missing values were replaced using 10-nearest neighbor imputation.

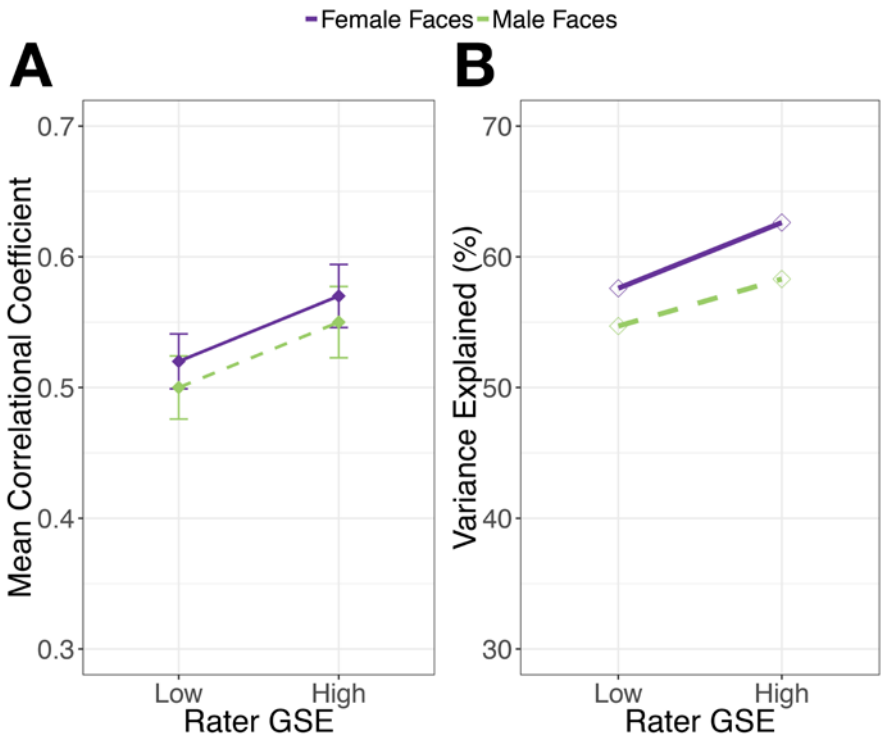

Figure S2. The level of intercorrelations across impressions (A) and the amount of variance in the impressions explained by valence (B) as a function of the raters' GSE score in Study 1b. Each data point (A: the absolute value of correlational coefficients between all impression rating pairs, B: the amount of variance explained by PC1 (valence) in the PCA of ratings per gender) was calculated from two rater subgroups ($n_{high\text{-}GSE}$ = 235, $n_{low\text{-}GSE}$ = 234). We divided the participants into high- and low-GSE raters for each trait, using the median split per trait (range of median = [112.5,126.0] across traits). For the raters whose GSEs were exactly at the median per trait (n = 13), we categorized 7 of them with higher GSEs in the high-GSE group and 6 with lower GSEs in the low-GSE group, regardless of which trait they evaluated faces on. The intercorrelations and the variance explained were higher in female than in male impressions among both high- ($M_{|r|}$ = 0.57, $SD_{|r|}$ = 0.23 vs. $M_{|r|}$ = 0.55, $SD_{|r|}$ = 0.26, $\chi^2(91) = 799.53$, $P < .001$; 62.62 % vs. 58.30 %) and low-GSE raters ($M_{|r|}$ = 0.52, $SD_{|r|}$ = 0.20 vs. $M_{|r|}$ = 0.50, $SD_{|r|}$ = 0.23, $\chi^2(91) = 545.05$, $P < .001$; 57.59 % vs. 43.70 %). These results replicate the results in the main text: Figure 3 shows consistent results, with individual differences in the GSE level better preserved. The error bars denote ±SE. GSE = gender stereotype endorsement**.** PCA = principal component analysis.

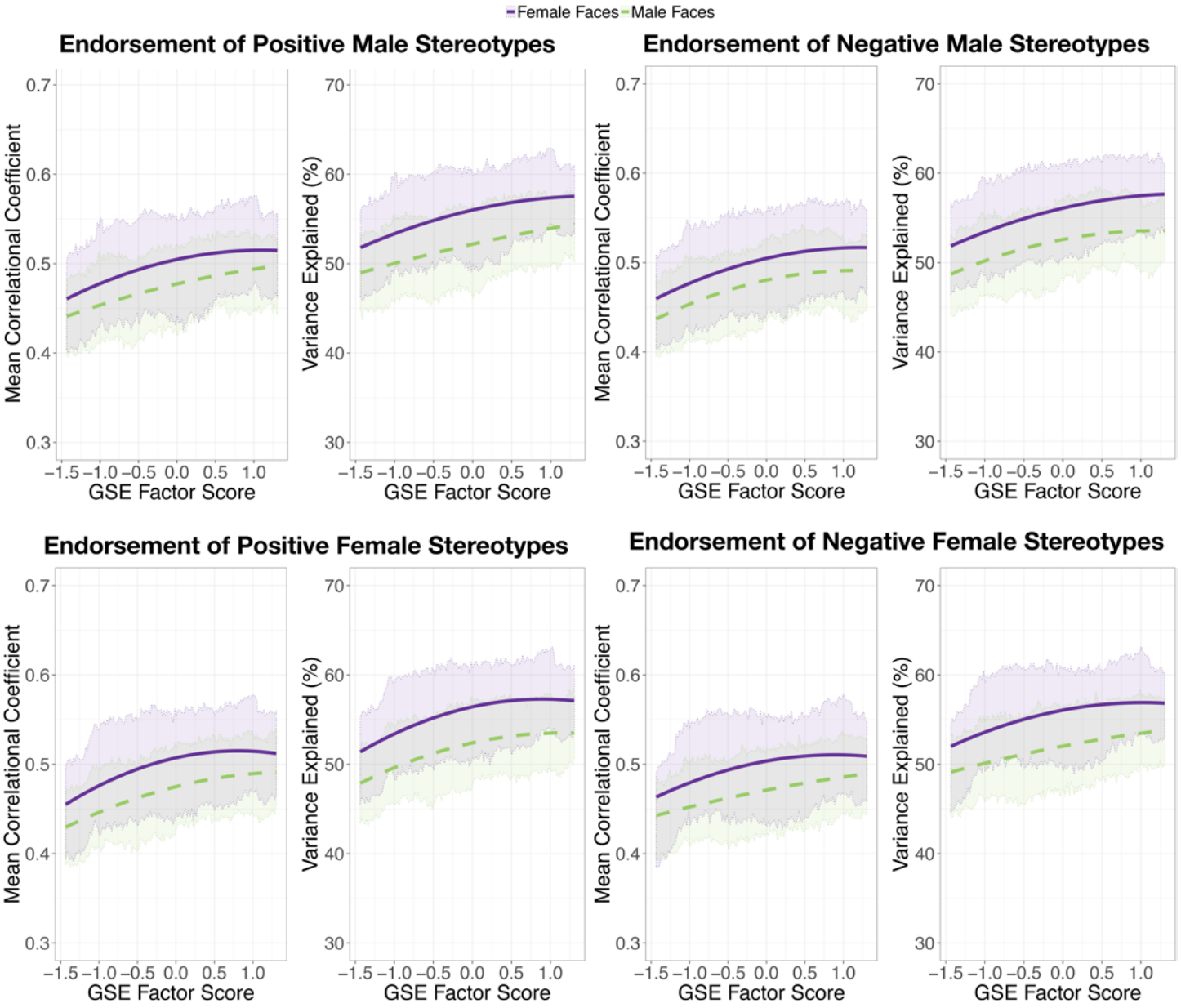

Figure S3. The level of intercorrelations across impressions (the left column in each subpanel) and the amount of variance in the impressions explained by valence (the right column in each subpanel) as a function of the raters' GSE factor scores in Study 1b. Each data point was calculated from a rater subgroup (n_{rater} = 10 per trait, n_{rater} = 140 in total per subgroup). Each subgroup was sampled from a sliding window on the rater GSE factors (n_{rater} ≥ 10 per trait), in which the *X* value is the middle point of the sliding window. GSE factors represent the degree to which raters endorsed gender stereotypes about either stereotypically male and positive (e.g., analytical; top left), male and negative (e.g., hostile; top right), female and positive (e.g., nurturing; bottom left), or female and negative traits (e.g., nagging; bottom right). The factor scores were derived from a four-solution confirmation factor analysis. The shaded regions show 95% CIs estimated from 1,000 bootstrapped replications per face gender for each factor score. The intercorrelations of face impressions (*t*s > 2.14, *P*s < .033) and the amount of variance explained by PC1 (*t*s > 7.69, *P*s < .001) were

significantly higher in female than in male impressions across every GSE factor score. See *Supplemental Text* for details. GSE = gender stereotype endorsement. CI = confidential interval. PC = principal component.

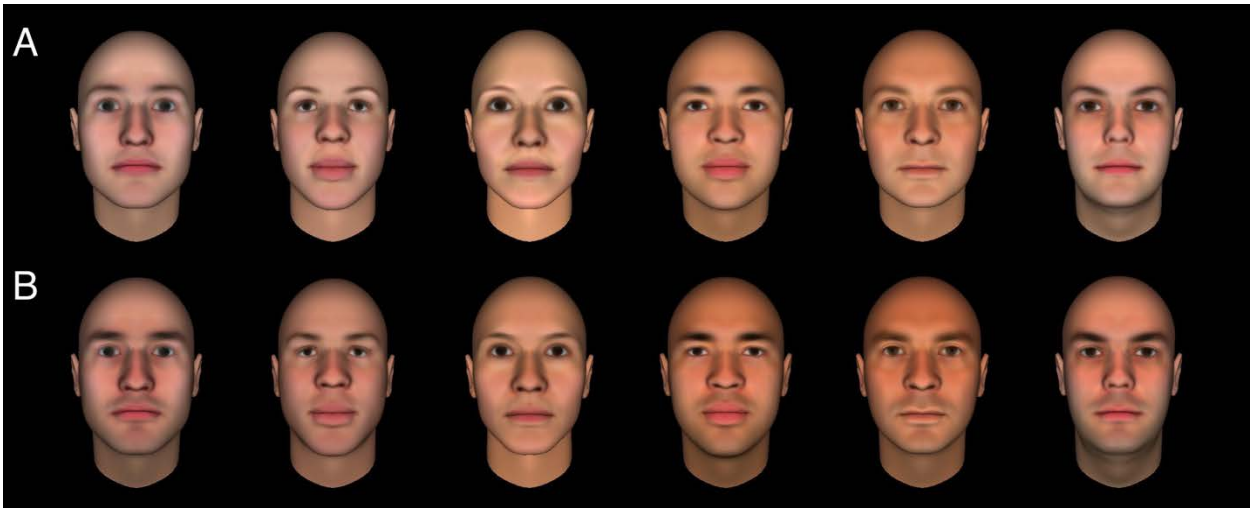

Figure S4. A sample of randomly generated synthetic female faces (A) and male faces (B) in Study 2. For each gender, 300 faces were generated as variations of the gender-specific average face. The face shape and face reflectance were varied randomly.

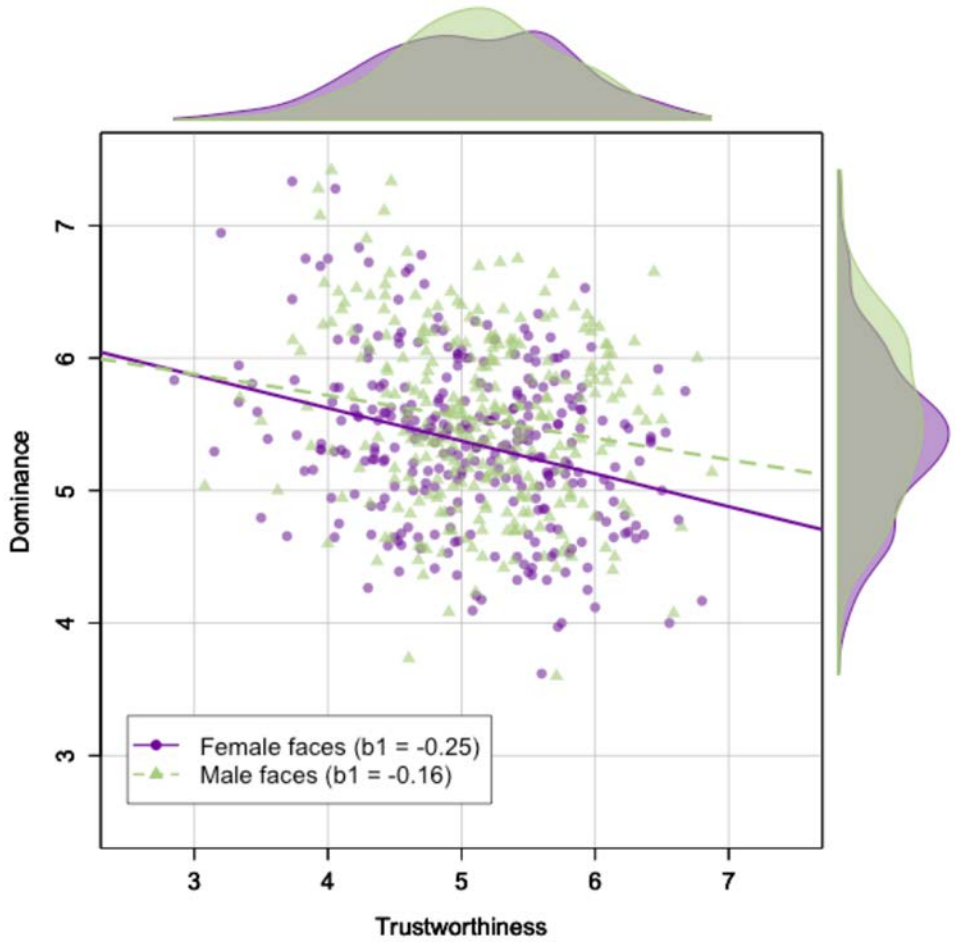

Figure S5. A scatterplot of the dominance and trustworthiness ratings of faces as a function of gender in Study 2. The density functions along the *X* and *Y* axes represent the distributions of male (green) and female faces (purple) for the trustworthiness and dominance ratings, respectively.

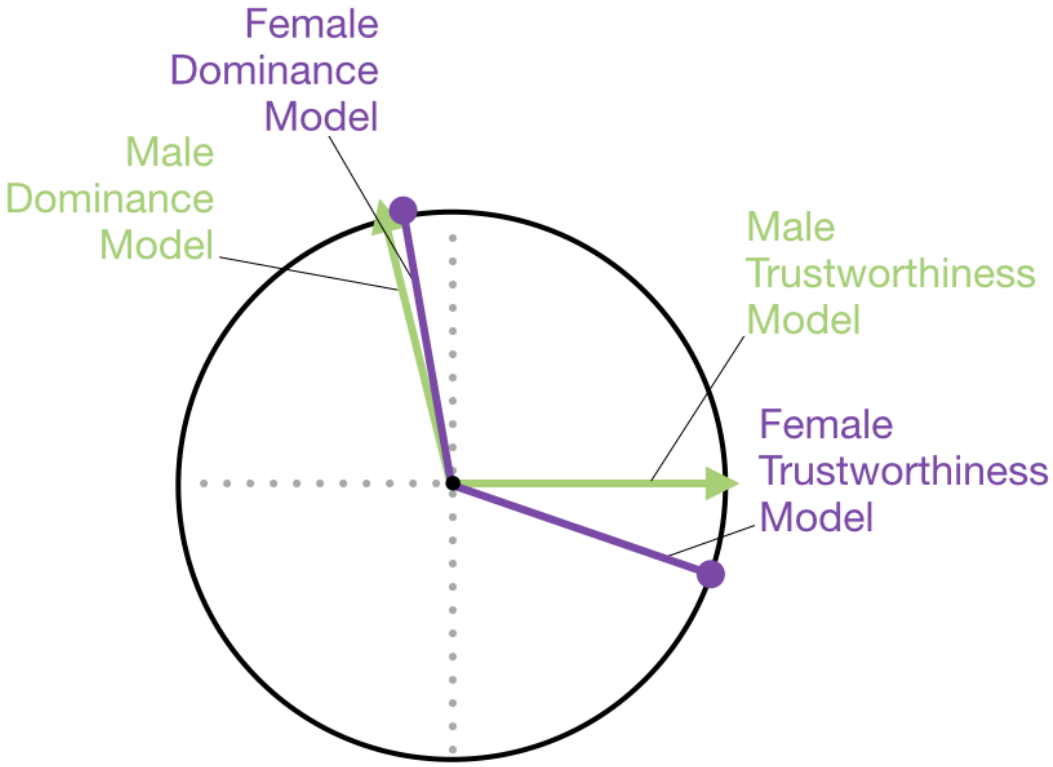

Figure S6. Similarities between gender-specific trustworthiness and dominance models. Similarities between models are represented with angles, e.g., 0 rad when $\rho = 1$, $\pi/2$ rad when $\rho = 0$, π when $\rho = -1$. Each model was built on the ratings of either only male faces (green line) or only female faces (purple line).

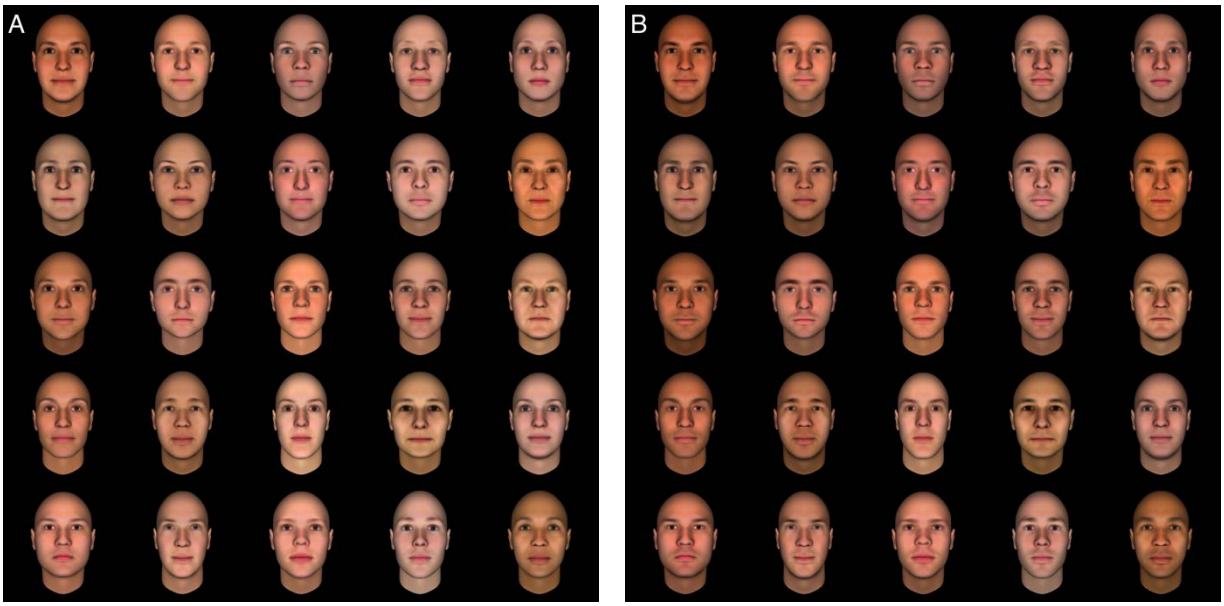

Figure S7. Twenty-five female (A) and twenty-five male (B) synthetic face identities used in the validation of the data-driven, computational models in Study 3a. These faces were randomly generated by a statistical face model with the constraint to be maximally distinctive from each other.

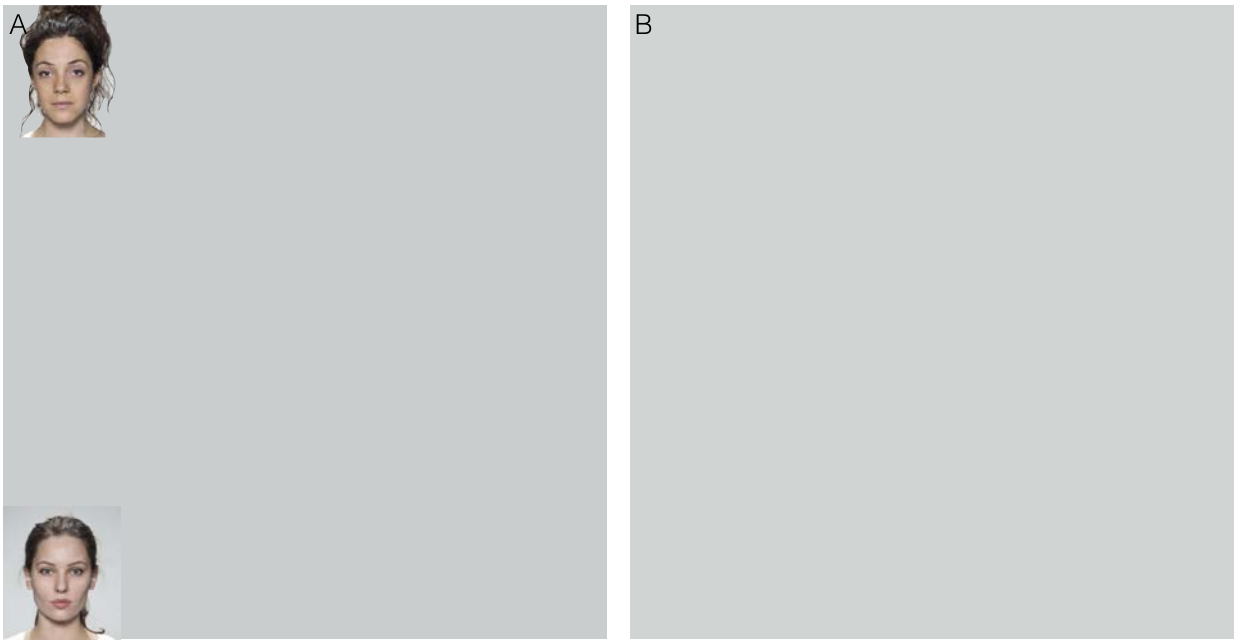

Figure S8. Twenty-five female (A) and twenty-five male (B) real-life face identities used in the validation of the data-driven, computational models in Study 3b. Images from *Face research lab London set*, by L. M. DeBruine and B. C. Jones, 2017. Retrieved from http://dx.doi.org/10.6084/m9.figshare.5047666

Table S1. Eigenvalues of PCs and the Amount of Variance Explained by the PCs in Study 1.

	Study 1a				Study 1b				Study 1c			
	Female Faces		Male Faces		Female Faces		Male Faces		Female Faces		Male Faces	
Component	λ	% Variance Explained	λ	% Variance Explained	λ	% Variance Explained	λ	% Variance Explained	λ	% Variance Explained	λ	% Variance Explained
1	**10.04**	**71.69%**	**8.18**	**58.40%**	**9.47**	**67.66%**	**8.67**	**61.94%**	**6.13**	**40.87%**	**4.74**	**31.60%**
2	**1.95**	**13.93%**	**3.38**	**24.16%**	**2.34**	**16.74%**	**3.45**	**24.65%**	**1.96**	**13.08%**	**3.13**	**20.84%**
3	0.83	5.91%	0.86	6.17%	0.58	4.15%	0.55	3.91%	**1.66**	**11.03%**	**1.94**	**12.91%**
4	0.39	2.78%	0.60	4.32%	0.39	2.77%	0.37	2.62%	**1.45**	**9.67%**	**1.30**	**8.65%**
5	0.17	1.22%	0.32	2.26%	0.28	2.03%	0.24	1.70%	0.99	6.61%	0.85	5.67%
6	0.15	1.05%	0.16	1.17%	0.20	1.43%	0.17	1.20%	0.76	5.05%	0.78	5.20%
7	0.12	0.85%	0.14	1.03%	0.16	1.11%	0.14	0.98%	0.57	3.83%	0.60	3.97%
8	0.10	0.71%	0.10	0.68%	0.15	1.05%	0.11	0.82%	0.46	3.06%	0.38	2.50%
9	0.08	0.55%	0.07	0.53%	0.13	0.92%	0.08	0.56%	0.29	1.91%	0.31	2.09%
10	0.05	0.36%	0.06	0.41%	0.10	0.69%	0.07	0.49%	0.21	1.39%	0.27	1.83%
11	0.04	0.32%	0.04	0.29%	0.08	0.59%	0.05	0.37%	0.15	0.97%	0.22	1.45%
12	0.03	0.23%	0.03	0.22%	0.05	0.36%	0.04	0.31%	0.14	0.91%	0.19	1.23%
13	0.03	0.21%	0.02	0.18%	0.05	0.34%	0.04	0.27%	0.11	0.75%	0.13	0.84%
14	0.03	0.18%	0.02	0.17%	0.02	0.17%	0.03	0.19%	0.07	0.45%	0.11	0.76%
15	(n/a)	(n/a)	(n/a)	(n/a)	(n/a)	(n/a)	(n/a)	(n/a)	0.06	0.40%	0.07	0.46%

Note. Boldface indicates eigenvalue > 1.00. PC = principal component.

Table S2. Factor Loadings of GSE Items from the CFA in Study 1b.

Trait Gender	Trait Valence	Trait Item	Factor 1	Factor 2	Factor 3	Factor 4
Stereotypical Male Traits	Positive	**dominant**	**.79**	.00	.00	.00
		competitive	**.68**	.00	.00	.00
		quantitative	**.53**	.00	.00	.00
		analytical	**.50**	.00	.00	.00
	Negative	**aggressive**	.00	**.75**	.00	.00
		hostile	.00	**.72**	.00	.00
		egotistical	.00	**.71**	.00	.00
		boastful	.00	**.70**	.00	.00
		arrogant	.00	**.67**	.00	.00
		cynical	.00	**.31**	.00	.00
Stereotypical Female Traits	Positive	**sensitive**	.00	.00	**.78**	.00
		nurturing	.00	.00	**.75**	.00
		artistic	.00	.00	.30	.00
		intuitive	.00	.00	.21	.00
	Negative	**emotional**	.00	.00	.00	**.80**
		nagging	.00	.00	.00	**.71**
		subordinate	.00	.00	.00	**.58**
		whiny	.00	.00	.00	**.58**
		servile	.00	.00	.00	**.56**
		gullible	.00	.00	.00	**.48**

Note. Boldface indicates factor loading > .3. Each question read "How do the average man and the average woman compare with each other on how [TRAIT TERM] they are?" GSE = Gender Stereotype Endorsement. CFA = Confirmatory Factor Analysis.

Table S3. Similarity between Face Impression Models in Studies 2–3.

	Male models		Female models	
Trait Model	Trust-worthiness	Dominance	Trust-worthiness	Dominance
Male Trustworthiness	-	-	-	-
Male Dominance	-.16	-	-	-
Female Trustworthiness	**.68**	**-.44**	-	-
Female Dominance	-.14	**.85**	**-.38**	-

Note. Numbers indicate pairwise Pearson coefficients between computational impression models (100 parameters each). Boldface indicates $P < .001$. Figure S5 visualizes the same information except for the bottom two rows in the table. The original trustworthiness and dominance models from Oosterhof & Todorov (2008) were built without taking into account face gender.

Table S4. Interrater Reliabilities of Ratings of Synthetic Faces Manipulated by Trustworthiness and Dominance Models in Study 3a.

	Trait model and original face gender	Cronbach's α based on face ratings
Face × Model Gender-Congruent Faces	Male faces manipulated with male trustworthiness model	.97
	Male faces manipulated with male dominance model	.96
	Female faces manipulated with female trustworthiness model	.96
	Female faces manipulated with female dominance model	.97
Face × Model Gender-Incongruent Faces	Male faces manipulated with female trustworthiness model	.96
	Male faces manipulated with female dominance model	.98
	Female faces manipulated with male trustworthiness model	.97
	Female faces manipulated with male dominance model	.96

Table S5. Interrater Reliabilities of Ratings of Real-Life Faces Manipulated by Trustworthiness and Dominance Models in Study 3b.

	Trait model and original face gender	Cronbach's α based on face ratings
Face × Model Gender-Congruent Faces	Male faces manipulated with male trustworthiness model	.88
	Male faces manipulated with male dominance model	.93
	Female faces manipulated with female trustworthiness model	.81
	Female faces manipulated with female dominance model	.92
Face × Model Gender-Incongruent Faces	Male faces manipulated with female trustworthiness model	.83
	Male faces manipulated with female dominance model	.91
	Female faces manipulated with male trustworthiness model	.85
	Female faces manipulated with male dominance model	.92

References

DeBruine, L. M., & Jones, B. C. (2017). Face Research Lab London Set. *Figshare*. http://doi.org/10.6084/m9.figshare.5047666

Dotsch, R., & Todorov, A. T. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*(5), 562–571. http://doi.org/10.1177/1948550611430272

Funk, F., Walker, M., & Todorov, A. T. (2016). Modelling perceptions of criminality and remorse from faces using a data-driven computational approach. *Cognition and Emotion*, *40*(5), 1–13.

Glick, P., & Fiske, S. T. (1996). The ambivalent sexism inventory: Differentiating hostile and benevolent sexism. *Journal of Personality and Social Psychology*, *70*(3), 491–512.

Glick, P., Fiske, S. T., Mladinic, A., Saiz, J. L., Abrams, D., Masser, B., et al. (2000). Beyond prejudice as simple antipathy: Hostile and benevolent sexism across cultures. *Journal of Personality and Social Psychology*, *79*(5), 763–775. http://doi.org/10.1037//0022-3514.79.5.763

Glick, P., Lameiras, M., Fiske, S. T., Eckes, T., Masser, B., Volpato, C., et al. (2004). Bad but bold: Ambivalent attitudes toward men predict gender inequality in 16 nations. *Journal of Personality and Social Psychology*, *86*(5), 713–728.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research*, *41*(17), 2261–2271.

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*(1), 269–297. http://doi.org/10.1146/annurev-psych-010416-044242

Jennrich, R. I. (1970). An asymptotic χ^2 test for the equality of two correlation matrices. *Journal*

of the American Statistical Association, *65*(330), 904–912. http://doi.org/10.1080/01621459.1970.10481133

Mangini, M., & Biederman, I. (2004). Making the ineffable explicit: estimating the information employed for face classifications. *Cognitive Science*, *28*(2), 209–226. http://doi.org/10.1016/j.cogsci.2003.11.004

Oh, D., Buck, E. A., & Todorov, A. T. (2019). Revealing hidden gender biases in competence impressions from faces. *Psychological Science*, 30(1), 65–79. http://doi.org/10.1177/0956797618813092

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. http://doi.org/10.1037/1082-989X.8.4.434

Oosterhof, N. N., & Todorov, A. T. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092. http://doi.org/10.1073/pnas.0805664105

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118. http://doi.org/10.1016/j.cognition.2012.12.001

Todorov, A. T., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE Signal Processing Magazine*, *28*(2), 117–122. http://doi.org/10.1109/MSP.2010.940006

Todorov, A. T., Dotsch, R., Porter, J. M., & Oosterhof, N. N. (2013). Validation of data-driven computational models of social perception of faces. *Emotion*, *13*(4), 724–738. http://doi.org/10.1037/a0032335

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

http://doi.org/10.1111/j.1751-9004.2011.00389.x

Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, *9*(11), 1–13. http://doi.org/10.1167/9.11.12

Walker, M., & Vetter, T. (2016). Changing the personality of a face: Perceived Big Two and Big Five personality factors modeled in real photographs. *Journal of Personality and Social Psychology*, *110*(4), 609–624.