GENDER BIASES IN FACIAL IMPRESSIONS

DongWon Oh

A DISSERTATION
PRESENTED TO THE FACULTY
OF PRINCETON UNIVERSITY
IN CANDIDACY FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY
THE DEPARTMENT OF
PSYCHOLOGY
[Adviser: Alexander Todorov]

June 2018
Abstract

First impressions from faces are consequential, shaping important social outcomes. In this thesis, based on converging evidence from empirical studies and computational models, I show that women are at a disadvantage because of gender biases in impressions. First, impressions of women are less differentiated than impressions of men and more highly valence-laden (Part I). Specifically, impressions of social traits (e.g., trustworthiness, dominance) are more highly intercorrelated for women than for men. Second, computational models of first impressions show that although people use similar facial information when forming impressions of women and men, they interpret this information differently (Part II). For instance, while masculinity cues tend to contribute to positive impressions of men, they tend to contribute to negative impressions of women. Third, this is particularly salient in the case of impressions of competence (Part III). Specifically, when faces are manipulated to look competent but not attractive, competence impressions correlate highly with multiple measures of masculinity. The current work shows that (1) impressions of women are more simplified than those of men; with positive impressions of women being contingent on the typicality of their looks and (2) women are perceived as less competent than men; with competence impressions of women being contingent on their attractiveness. The current work highlights the importance of social categorization in first impressions and reveals a major obstacle to gender equality and social justice at large.
Acknowledgement

I would like to thank people who helped me begin and complete this fruitful journey of doctoral training and dissertation writing. These individuals’ support has been integral to my academic work and general well-being in Princeton.

First and foremost, I am grateful to my advisor, Alexander Todorov, who has given me the opportunity to research fascinating questions with him and has taught me the scientific attitude, methods, communication, and more – he guided me through every step of the way. I would also like to thank my other collaborators on the project of gender biases facial impressions: Ron Dotsch, Jenny Porter, and Nora Buck – this dissertation could never have been started or finished without their hard work. I also thank my proposal and dissertation committee members, who provided me with valuable feedback and questions, thereby helping me improve my work: Alin Coman, Eldar Shafir, Diana Tamir, Johannes Haushofer, and Yael Niv.

I thank the members of Social Perception Lab, whom I always relied on for scientific and personal advices: Aaron Kurosu, Joel Martinez, Justin Junge, Mai Nguyen, Gandalf Nicolas, and Brandon Labbree. I also thank previous members of Social Perception Lab, who taught me research methods and supported me in other ways: Shuo Wang, Peter Mende-Siedlecki, Virginia Falvello, and Jill Swencionis. I thank my advisor/collaborators on other projects: Dan Osherson, Eldar Shafir, Ajua Duker, Kimberly Solomon, Rae Drach, Joseph Avery, Alex Koch, among others. I thank the Psychology Dept staff: Sami Mezger, Keisha Craig, RoseMarie Stevenson, among others.

I cannot thank enough my friends, who shared many good times with me and helped me endure hard times: Yeon Soon Shin, Alex Kustov, Andrea Placidi, Alice Yoon,
Chaeseung Lim, Jeongseok Lee, Sebastian Musslick, Jamal Williams, Tasha Holden, Clare Choi, Jane Keung, Nayeon Kim, Bianca Dumitrascu, Jon Berliner, Sam McDougle, Angela Radulescu, Abby Novick, Michael Lloyd, Robin Gomila, Olga Lositsky, Pavlos Kollias, Jin Cheong, Yonghyun Song, Andrew Kim, Joonsuk Park, Junwoo Son, Boram Lee, Hongmi Lee, Ghootae Kim, Su Keun Jeong, Leo Park, Jung Won Han, Sungchul Park, Kyungbo Kim, Chul Moon, Geun Ho Yoo, Joon Hyun Ro, Jong Woo Yoon, Moon Sung Choi, Yoon-Jeong Yang, Jungwoo Kim, BG Kim, YooJeong Kim, and Taejun Kim, among many, many others.

I thank Clinical Neuroscience Lab members, who always inspire me academically and otherwise: Hanbyul Lee, Eunha Jeong, Solji Bae, Hyun Song, Hankyung Lee, Eunhee Park, Hyeon Park, Hoyoung Kim, Minyoung Shin, Jaeik Kim, Haeran Kim, Minue Kim, Youngbin Kwak, and especially Jeanyung Chey, my former advisor.

I thank my co-instructors, who taught me how to teach: Justin Junge, Yael Niv, Liz Gould, Asif Ghazanfar, Joel Cooper, and the staff and faculty of NJ-STEP (New Jersey Scholarship and Transformative Education in Prisons Consortium).

Graduating makes me look back and think about how all this started – I thank scholars/teachers at my prior institutions who believed in my potential as a researcher and introduced me into the amazing world of scientific research: Choongkil Lee, Sang-Hun Lee, Randolph Blake, Byoung-Tak Zhang, Sungryong Koh, and the researchers/staff at Korea Foundation for Advanced Studies, among others.

Lastly, I would like to thank my family: my grandfather, my father, and my mother, who always trusted me in everything; and Panyagal Charmroenpucks, my wonderfully trustworthy and dominant – therefore, competent – soon-to-be wife.
Table of Contents

Abstract iii

Acknowledgements iv

Contents vi

List of Figures viii

Introduction 10

Part I: Gender Biases in Interrelationship of Impressions 13

Study 1a 19

Study 1b 24

Study 1c 33

Part I: Interim Conclusion 35

Part II: Gender Biases in Models of Impressions 38

Study 2 39

Study 3a 46

Study 3b 51

Part II: Interim Conclusion 58

Part III: Gender Biases in Impressions of Competence 62

Study 4a 65

Study 4b 72

Study 5 76

Study 6 79
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 7</td>
<td>83</td>
</tr>
<tr>
<td>Part III: Interim Conclusion</td>
<td>89</td>
</tr>
<tr>
<td>Conclusion</td>
<td>93</td>
</tr>
<tr>
<td>References</td>
<td>96</td>
</tr>
<tr>
<td>Appendices</td>
<td>105</td>
</tr>
</tbody>
</table>
## List of Figures

Figure 1. The level of the intercorrelations of trait impressions of men and women. 22

Figure 2. Amount of the variance explained by the first (PC1) and the second component (PC2) derived from a principal component analysis (PCA) of trait ratings of faces. 23

Figure 3. The level of intercorrelations among impressions (A) and the amount of variance in the impressions explained by PC1 (B) as a function of the raters’ gender stereotype endorsement (GSE) level. 31

Figure 4. Models of female (top) and male (bottom) trustworthiness impression applied to a sample male face. 44

Figure 5. Models of female (top) and male (bottom) trustworthiness impression applied to a sample female face. 44

Figure 6. Models of female (top) and male (bottom) dominance impression applied to a sample male face. 45

Figure 7. Models of female (top) and male (bottom) dominance impression applied to a sample female face. 45

Figure 8. Validation of models of trustworthiness (top) and dominance (bottom) with computer-generated male (left) and female faces (right). 49

Figure 9. Models of female (top) and male (bottom) trustworthiness impression applied to a sample male face. 54

Figure 10. Models of female (top) and male (bottom) trustworthiness impression applied to a sample female face. 54
Figure 11. Models of female (top) and male (bottom) dominance impression applied to a sample male face.

Figure 12. Models of female (top) and male (bottom) dominance impression applied to a sample female face.

Figure 13. Validation of models of trustworthiness (top) and dominance (bottom) with real-life male (left) and female face images (right).

Figure 14. A face manipulated by the competence model (top), the attractiveness model (middle), and the difference[comp-attr] model (bottom).

Figure 15 The mean impression rating of competence (left) and attractiveness (right) as a function of the competence model (top) and the difference[comp-attr] model (bottom) manipulation (Studies 4a – 4b).

Figure 16. The mean impression rating of confidence (left) and masculinity (right) as a function of the difference[comp-attr] model manipulation (Study 6).

Figure 17. The mean proportion of “male” responses as a function of the level of the difference[comp-attr] model (black line) and the competence model (grey line) manipulation.

Figure 18. A sample of real-life face images manipulated by the difference[comp-attr] model.

Figure 19. The mean competence ratings of real-life male (left) and female faces (right) as a function of the difference[comp-attr] model manipulation (Study 7).
Introduction

“Look like a girl, act like a lady, think like a man, work like a boss

#HappyWomensDay”

– Bic Facebook ad (March 8, 2015)

Gender is central to one’s identity. The difference between genders can be so exaggerated that each gender is expected to have a certain set of unique characteristics, such as distinct social traits or cognitive abilities (Ellemers, 2018; Rudman & Glick, 2012), and those characteristics are thought to be shared by all members of the gender (Prentice & Miller, 2006). This category-based way of thinking neglects the huge overlap in social and cognitive characteristics between genders (Ellemers, 2018; Hyde, 2014). As a result, the way we treat each other varies significantly depending on our perceptions of gender, often leading to unfair treatment of women in various settings. The current thesis is concerned with the moment at which gender biases are first introduced: forming first impressions from facial appearance. Understanding gender difference in facial impressions is important, because these impressions are formed rapidly and effortlessly, and they are consequential.

People effortlessly attribute social traits, such as trustworthiness and dominance, to others based on their facial appearance (Todorov, 2017; Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). These impressions of traits affect various important real-world

---

1 The post was soon removed when met with criticism and contempt on the Internet. The company eventually issued an apology. The phrase in the ad exemplifies numerous social traits that are considered desirable in women. I am unaware of any male equivalent of these traits, not with the same level of normative strength.
outcomes, which range from voting behavior (Antonakis & Dalgas, 2009; Lenz & Lawson, 2011; Little, Burriss, Jones, & Roberts, 2007; Olivola & Todorov, 2010; Todorov, Mandisodza, Goren, & Hall, 2005), to court decisions (Blair, Judd, & Chapleau, 2004; Eberhardt, Davies, Purdie-Vaughns, & Johnson, 2006; Wilson & Rule, 2015; Zebrowitz & McDonald, 1991), and to mating choices (Cooper, Dunne, Furey, & O'Doherty, 2012; Little, Burt, & Perrett, 2006). Importantly, when forming impressions of others, we have a strong tendency to categorize them, especially along gender lines (Fiske & Neuberg, 1990). Thus, it is crucial to understand how gender categorization affects the process of impression formation. In the present dissertation, I address this issue, using behavioral experiments, data-driven computational modeling, dimensionality reduction techniques (e.g., principal component analysis, or PCA), and face morphing. The dissertation consists of three parts. In Parts I and II, I investigate whether the structures of impressions of men and women are different. In Part III, I investigate the visual ingredients of a specific consequential impression – competence – and whether these ingredients favor men over women.

Specifically, in Part I, I investigate how impressions from faces (i.e., trustworthiness, attractiveness, competence, and so on) are related to each other and whether these relations are different for men and women. I show that impressions of women are less differentiated and more highly driven by overall positivity/negativity of impressions than those of men. In Part II, using data-driven computational methods, I test for perceptual differences in the impressions of men and women. I find that people use essentially the same visual information when forming impressions of men and women. This finding suggests that gender categorization changes how people interpret the same
information for men and women. For example, while masculinity cues tend to contribute to positive impressions of men, they tend to contribute to negative impressions of women. In Part III, I investigate whether an important, consequential impression – competence – is gender-biased. Impressions of competence are highly correlated with attractiveness. Using data-driven methods, I create a model of competence that controls for facial attractiveness. Controlling for attractiveness shows that people rely heavily on facial masculinity when forming impressions of competence. This bias favors men, because masculine-appearing women are evaluated negatively.

I hope that my thesis will add to the growing scientific literature on social cognition and gender biases. For the basic science, the current work reveals deleterious gender biases in the process of first impression formation, highlighting the significance of social categories in visual social perception. For the applied science, the current work hints at how to fight gender biases in social perception at the micro (e.g., how to effectively manage one’s impressions to counteract stereotypes) as well as macro levels (e.g., how to create a fair professional interview setting for candidates of all genders).
Part I: Gender Biases in Interrelationship of Impressions

An important way to understand how people form impressions of others is to investigate the correlations between individual impressions (e.g., the correlation between impressions of attractiveness and trustworthiness). Because the trait impressions are highly intercorrelated, by examining the relations among the perceived traits, one can succinctly describe the structure of face impression formation, predict impressions based on multiple facial features, and reveal the facial information used to form these impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov, Dotsch, Porter, & Oosterhof, 2013; Walker & Vetter, 2009; 2016).

A body of research has studied the structure underlying the relations between traits in social impressions (Fiske, Cuddy, Glick, & Xu, 2002; Imhoff & Koch, 2016; Koch, Imhoff, Unkelbach, & Alves, 2016; Oosterhof & Todorov, 2008; Osgood, Suci, & Tannenbaum, 1957; Rosenberg, Nelson, & Vivekananthan, 1968; Wiggins, 1979). In this line of work, one typically extracts a small number of components using a dimensionality reduction technique, such as PCA. Mathematically, each component is a linear combination of all impressions that can explain a large amount of variance in the impression ratings. Conceptually, each component is a latent variable that summarizes the individual impressions that have a high level of correlations with the component. In facial impressions, using this approach, one can reduce the impressions to two summary components: valence and physical power. These two components can be approximated by the intuitive judgments of trustworthiness and dominance of faces, respectively (Oosterhof & Todorov, 2008; Sutherland, Oldmeadow, & Young, 2016; Walker & Vetter, 2016). Importantly, the structure of the face-based impressions (i.e., what components
can summarize the face-based impressions) may vary across meaningful subcategories of faces, such as male and female faces. However, previous research on face impressions has ignored potential gender differences, implicitly assuming the same structure of impressions across genders. This assumption is largely inconsistent with both empirical evidence and theoretical reasoning.

Empirical data suggest that impressions from facial appearance are more highly correlated to each other for women than for men. Sutherland et al. (2015) found that trustworthiness and dominance impressions were negatively correlated for female faces, whereas the two impressions were not or only weakly correlated for male faces (Sutherland, Young, Mootz, & Oldmeadow, 2015). They also found that dominant female faces were more negatively evaluated (e.g., perceived as colder, sterner, and less trustworthy) compared to non-dominant female faces, non-dominant male faces, and dominant male faces. These findings are inconsistent with the existing model of face trait impressions, which assumes that the two summary dimensions of valence/trustworthiness and power/dominance are orthogonal to each other. Given the high correlations of these two trait impressions with other trait impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013), it is likely that face-based trait impressions are more highly intercorrelated for women than for men, implying less differentiated face impressions of women.

This idea aligns well with the rich literature of gender stereotypes. People expect women to be more submissive, dependent, and gentle, and less dominant, independent, and aggressive than men (Bem, 1974; Prentice & Carranza, 2002; Rudman & Glick, 2012; Spence, Helmreich, & Holahan, 1979; Spence, Helmreich, & Stapp, 1975; Vogel,
Broverman, Clarkson, & Rosenkrantz, 1972). These beliefs are widely held across cultures (Glick et al., 2000; Williams & Best, 1990b; Williams, Satterwhite, & Best, 1999) and difficult to change (Prentice & Carranza, 2004). Although people have gender-associated expectations for both men and women, a larger number of social traits are considered more typical or desirable for women than for men (Prentice & Carranza, 2002), including valence-related traits, such as kindness, friendliness, and niceness (Heilman, 2001; Rudman, 1998; Rudman & Glick, 2001). Relatedly, women are evaluated positively to the extent that they conform to the stereotypes associated with them, the boundaries of which are narrower compared to those associated with men (Glick et al., 2000; Glick & Fiske, 1996). If these principles apply to face-based impression formation too, women whose looks are inconsistent with the stereotypes would be more likely to be perceived as less typical and desirable compared to men whose looks are inconsistent with the stereotypes. Thus, in impressions of women, more social traits should be correlated with valence compared to impressions of men, as shown by the negative correlation between perceived facial dominance and perceived trustworthiness/valence for women (Sutherland et al., 2015). Thus, I expect less differentiated face impressions for female than for male faces. Specifically, I expect a higher level of correlation between all impression pairs for female than for male faces. In other words, at the face level, different impressions (e.g., attractiveness, competence) would covary more strongly for female than for male faces.

Using dimensionality reduction and computational modeling of multiple datasets, the current thesis goes beyond the examination of the relations between two face impressions (e.g., dominance and valence), and examines the general principles behind
the gender difference in the face-based impressions. In Part I, by examining the degree of intercorrelations between the ratings of face impressions, I investigate whether and to what extent female face impressions are less differentiated than male face impressions. Impressions of dominance and valence are more strongly negatively correlated for women than for men (Sutherland et al., 2015). However, it is yet to be tested whether this stronger correlation in female than in male impressions is true for all impressions. Across three different face datasets, I found that female impressions are, indeed, less differentiated (Studies 1a – 1c). Further, I tested whether the degree of differentiation is related to how much individual raters endorsed gender stereotypes (Study 2). Further, I test whether the degree of differentiation is related to a rater characteristic related to gender stereotypes, namely, how much raters endorse gender stereotypes.

In Part II, by building separate data-driven computational models of impressions (Oosterhof & Todorov, 2008) of male and female faces, I investigate the mechanisms that underlie the less differentiated impressions of women. Specifically, the gender-related difference in impression differentiation can stem from either (1) the same visual information used differently across face genders to form an impression or (2) different visual information used to form the impression. For facial attractiveness, for example, the same visual information is used across genders but has the opposite evaluative outcome, in which masculine facial color (e.g., darker skin) increases the attractiveness of men but decreases the attractiveness of women (Said & Todorov, 2011). This supports the possibility that the gender difference in impression differentiation stems from the same information used differently across genders. However, it is yet to be tested whether this would generalize to key face impressions, such as trustworthiness and dominance.
A data-driven face model of a trait impression (e.g., trustworthiness) represents visual information that people use to form this impression (Dotsch & Todorov, 2012; Jack & Schyns, 2017; Todorov, Dotsch, Wigboldus, & Said, 2011). Specifically, an impression model is a description of facial features that covary with a trait impression of faces, derived from people’s ratings of faces (e.g., how dominant each face looks; an impression model will be operationally defined once the data-driven modeling approach is introduced in Studies 2 and 3). If people use the same information when forming impressions of male and female faces (e.g., masculine features to infer dominance), then the models of male and female impressions should be similar. Such a result would imply that different evaluative process causes differences in impression differentiation between male and female faces due to gender categorization. Categorizing a face as female, for example, would lead to stronger correlations between impressions that reflect female gender stereotypes. On the other hand, if people use different information when forming impressions of male and female faces, then models of male and female impressions should be different.

One can test these possibilities by (a) looking directly at the similarity of the male and female models and (b) cross-validating the models of impressions on novel male and female faces. To the extent that the models for male and female faces are similar, the models would be highly correlated, and their effects on impressions would be similar, regardless of whether they are applied to male or female faces. For example, impressions of the trustworthiness of a novel female face would be similar regardless of whether the face is manipulated by a male or a female model of trustworthiness. In contrast, even if the models for male and female faces are highly correlated, to the extent that they are
based on different information, they would have different effects on impressions depending on whether they are applied to a male or a female face (e.g., impressions of the trustworthiness of a novel female face would be more successfully manipulated by a female than by a male model of trustworthiness).
Study 1a: Introduction

In Studies 1 – 2, I assessed the gender difference in the level of impression differentiation by comparing two measures: (a) the average degree of correlation across all impression pairs in each gender and (b) the degree to which impressions are explained by valence of impressions. I expected less differentiated face impressions for women than for men expressed through (a) a stronger correlation between trait impressions (e.g., trustworthiness and dominance) for female than for male faces and (b) a larger variance explained by the first principal component (PC1), which typically captures the valence of impressions, for female than for male faces.

In Study 1a, I conducted the abovementioned two analyses on preexisting rating datasets of male and female face images (Oosterhof & Todorov, 2008). In the original work, Oosterhof and Todorov (2008) conducted a PCA on the rating dataset without considering face gender, and found the first two principal components accounted for >80% of the variance, which could be interpreted as valence and physical power, respectively.

Study 1a: Methods

Participants. Three-hundred-and-one Princeton University students were recruited by Oosterhof and Todorov (2008), and participated in the trait rating experiments for partial course credit or cash.

Stimuli. Naturalistic face photos with direct gaze and neutral expressions were used. The individuals were amateur actors with no facial hair, earrings, eyeglasses, or visible make-up, all wearing grey T-shirts. Sixty-six photos (33 males, 33 females) of White actors between the ages of 20–30 of were used (Lundqvist, Flykt, & Arne, 1998).
Procedure. Participants rated 33 male and 33 female face photos on 14 social traits: how aggressive, attractive, caring, confident, dominant, emotionally stable, intelligent, mean, responsible, sociable, threatening, trustworthy, unhappy, or weird each individual looked. These traits were selected due to their empirical and theoretical importance: the traits (except for dominance) explained about 68% of unconstrained, spontaneous person descriptions from face images (Oosterhof & Todorov, 2008). Dominance was included because of its importance in personality perception (Wiggins, 1979) and face evaluation (Oosterhof & Todorov, 2008). Each participant rated all 66 faces for only one trait impression.

To collect the face impression ratings, different groups of participants were assigned to form impressions of all 66 faces on a single social trait \( n_{\text{min}} = 18 \). Participants were told that the study was about first impressions and were encouraged to rely on their “gut feeling.” The faces were presented and rated for three times in separate blocks to reduce the measurement error for each participant’s answers. The average of each participant’s ratings of the faces served as the measure of their evaluation on the respective trait. This procedure also allows for screening out unreliable raters: those who show zero or negative test-retest within-rater reliability as calculated between ratings in different blocks. Each face image was presented at the center of the screen with a question above the face (“How [trait term] is this person?”) and a response scale below the face (“1 Not at all [trait term] - 9 Extremely [trait term]”). Each face was visible until the participant responded, the intertrial interval (ITI) was 1,000 ms, and the order of faces was randomized. All 14 trait ratings showed moderate to high intrarater agreement \( r_{\text{min}} \).
= .26) and interrater reliability ($\alpha_{\text{min}} = .90$). To obtain impression measures for each face, the ratings on the 14 social traits were averaged across raters.

I conducted two main analyses to test for gender differences in the level of differentiations in impressions. In the first analysis, I calculated the extent to which social traits are correlated to one another in each gender, and compared the level of correlations between ratings of male and female faces. Specifically, I computed the correlations among trait ratings separately for male and female faces (see Appendix A for the correlational matrices) – higher absolute values of the correlational coefficients implicate a higher dependency between perceptions of traits – and then using a paired $t$-test, I compared the absolute values of the coefficients for male and female face ratings. Correlational coefficients between the ratings of every possible trait pair among all 14 traits ($C(14,2) = 91$ pairs) were calculated on the face image level for each gender. In the second analysis, I $z$-transformed the average trait ratings of faces within each trait, and then subjected them to a PCA for each gender. To test whether raters showed less differentiated impressions for female than for male faces, I compared the amount of the variance explained by PC1 for each gender: PC1 loaded highly on valenced impression ratings (e.g., trustworthiness), and a higher value of PC1 would mean bigger dependency of impressions on valence.

**Study 1a: Results and Discussion**

For both genders, PC1 was highly loaded on by all positive (e.g., trustworthy, responsible) and negative (e.g., threatening, weird) traits (see Appendix B for the PCA loading strength). This is consistent with previous models of face impressions, in which
the first component is summarized as valence (Oosterhof & Todorov, 2008) or approachability (Sutherland et al., 2015). Importantly, new to the current data, female face impressions were less differentiated and more valence-laden than male face impressions (Fig. 1): the absolute value of the correlational coefficients were bigger for female than for male face ratings ($t(90) = 8.64, P < .001$), indicating a higher level of dependency between impressions in female face impressions. This is consistent with a visual inspection of the PCA solutions: the trait loadings on PC1 have bigger absolute values for female than for male faces (Appendix B). Correspondingly, the amount of variance explained by PC1, a proxy for valence or approachability, was larger for female than for male ratings (71.69% vs. 58.40%; Fig. 2), indicating a higher level of dependency of impressions on overall positivity/negativity in female face impressions.

Fig. 1. The level of the intercorrelations of trait impressions of men and women. In each study, correlational analyses were conducted at the face level between impression ratings separately for male and female images, and absolute values of the coefficients were compared across genders. Each dot corresponds to an impression pair (e.g., the absolute value of the coefficient of the correlation between the
“threatening” and the “unhappy” impression rating). The contour of the violin plots represents the distribution of the values in each face gender in each study. The lower and upper hinges of each white box correspond to the 25th and 75th percentiles of the value, respectively. The black bar in each box denotes the median of the value. A higher $y$ value represents a lower level of differentiation (or a higher level of intercorrelation) between trait impressions. Across studies, the impressions of women are significantly more highly intercorrelated than the impressions of men ($P < .001$, $P < .01$, and $P < .05$, respectively).

**Fig. 2.** Amount of the variance explained by the first (PC1) and the second component (PC2) derived from a principal component analysis (PCA) of trait ratings of faces. In each study, a PCA was conducted separately for the impressions of male and female images. A higher $y$ value on PC1 represents a stronger relationship between the valence component of impressions and specific impressions. Across studies, impressions of women are more highly valence-laden than the impressions of men.
Study 1b: Introduction

Study 1a revealed that impressions of women are less differentiated and more valence-laden compared to impressions of men. What is the psychological basis of this gender difference in impressions? The objective of Study 1b was to replicate the findings of Study 1a and to test whether the differences in impressions can be explained by the gender stereotypes held by perceivers. Specifically, I expected that the more strongly a perceiver endorses conventional beliefs about genders, the more likely their ratings will show (a) less differentiated impressions of women than impressions of men and (b) more highly valence-laden impressions of women than impressions of men. Perceivers who endorse gender stereotypes will rate counterstereotypical faces more negatively, more so when the target face is a female than a male due to a larger number of and stronger prescriptive norms for women than for men. This will lead to more simplified and valence-laden impressions of women than of men in these perceivers. To test this idea, I collected impression ratings of male and female faces and measured participants’ gender stereotype endorsement.

Additionally, I recorded participants’ gender to test whether the participant gender is related to face-gender differences in impressions. The dataset used in Study 1a did not contain information regarding the participants’ gender; as a result, Study 1a could not investigate the effect of the participant gender on face-based impressions. A possibility is that male raters will show more simplified, valence-laden impressions of women than of men, given prior studies showing a higher level of endorsement of gender stereotypes in male than female raters (Glick & Fiske, 1996; Swim, Aikin, Hall, & Hunter, 1995; Williams & Best, 1990). Another possibility is that male and female raters will form
similar impressions across face genders, given prior studies showing no effect of participant gender on gender-stereotyping (Costrich, Feinstein, Kidder, Marecek, & Pascale, 1975; Deaux & Lewis, 1984; Eagly & Steffen, 1984; Goldberg, 1968; Hagen & Kahn, 1975; Moss-Racusin, Dovidio, Brescoll, Graham, & Handelsman, 2012). A final possibility is that female raters will show more simplified, valence-laden impressions of women than of men, given prior studies showing a higher level of gender-stereotyping (e.g., negative evaluation of women with counterstereotypical traits) in female than in male raters (Garcia-Retamero & López-Zafra, 2006; Goldberg, 1968; Parks-Stamm, Heilman, & Hearns, 2007; Rudman, 1998).

**Study 1b: Methods**

*Participants.* Five-hundred-thirty-six online participants (278 males, 258 females) participated through Amazon MTurk for monetary reward. Required participant number for each trait was estimated with the interrater reliabilities from Study 1a so that Cronbach’s $\alpha$ of the ratings would reach .90 for both male and female faces.

*Stimuli.* The naturalistic face photos of 33 men and 33 women used in Study 1a were used again (Lundqvist et al., 1998).

*Procedure.* Participants rated 33 male and 33 female face photos on the 14 social traits that were rated in Study 1a. As in Study 1a, different groups of participants were assigned to form impressions of all 66 faces on a single social trait ($n_{\text{min}} = 11$). Participants saw identical instructions with Study 1a. The faces were presented and rated twice in separate blocks to reduce the measurement error for each participant’s answers. The average of each participant’s ratings of the faces served as the measure of their evaluation on the
respective trait. As in Study 1a, I calculated each rater’s intra-rater reliability by correlating their ratings from different blocks. The ratings from participants with zero or negative reliability were excluded, which resulted in 469 participants’ responses.

Each face image was presented at the center of the screen with a question above the face (“How [trait term] is this person?”) and a response scale below the face (“1 Not at all [trait term] - 9 Extremely [trait term]”). Each face was visible until the participant responded, the ITI was 1,000 ms, and the order of faces was randomized. All 14 trait ratings showed moderate to high interrater agreement ($r_{\text{min}} = .32$) and interrater reliability ($\alpha_{\text{min}} = .81$). To obtain trait measures for each face, I averaged the ratings per face across raters for each trait. I conducted the same two main analyses as in Study 1a (i.e., the correlational analyses and the PCAs) to test for gender differences in the level of differentiations in impressions. First, I compared the degree of the intercorrelation across impressions between genders. Correlational coefficients between the ratings of every possible trait pair among all 14 traits ($C(14,2) = 91$ pairs) were calculated on the face image level for each gender. Second, I compared the amount of variance explained by PC1, a proxy of valence or approachability.

*Preliminary analyses.* To test for the effects of the raters’ gender and their gender stereotype endorsement, at the end of the study participants were asked to report their gender and to complete a questionnaire regarding gender stereotype endorsement (GSE) that measured the extent to which they agreed with conventional gender stereotypes (Cundiff & Vescio, 2016; Eagly & Mladinic, 1989). Each question read as follows: “How do the average man and the average woman compare with each other on how [trait term] they are?” The trait terms were 20 words describing traits either considered
stereotypically female and positive (e.g., nurturing), female and negative (e.g., whiny), male and positive (e.g., competitive), or male and negative (e.g., egotistical; see Appendix C for the list). The valence and the gender stereotypicality of these words had previously been validated (Cundiff & Vescio, 2016; Eagly & Mladinic, 1989; Spence et al., 1979).

Responses to every question in the questionnaire were significantly different from the middle score (5 in the range of [1,9]) in the stereotype-consistent direction ($t > 3.87$, $P < .001$; see Appendix D for the distribution of the responses to each question). The responses from participants with no variance were excluded (i.e. participant who answered all questions with a single value; $N = 17$), which resulted in responses of 452 raters. Each trait impression condition had $\geq$21 raters. Some responses were reverse-coded so that a larger value in each response always meant a higher level of GSE ([1,9] in each). Missing values were replaced with the average response to each question. The GSE questionnaire showed moderately high internal consistency across questions ($\alpha = .88$).

To create an index that represents each rater’s GSE level, I selected questions that contributed to a large degree to the total score of GSE. An item-whole correlation was calculated for each trait question (each item) as a measure of inter-item reliability: the correlation between each trait question and the whole questionnaire composed of the remaining questions when that question was left out. Some questions contributed more than others to the whole score (Appendix C). Based on the assumption that responses to all questions shared a single, identical latent variable (i.e., a rater’s GSE level), I selected questions with the highest inter-item reliability, specifically $>.6$ item-whole correlation.
Six questions survived the criterion: 3 of them were about traits stereotypically related to men (\textit{“competitive”}, \textit{“dominant”}, \textit{“aggressive”}) and the other 3 to women (\textit{“nurturing”}, \textit{“sensitive”}, \textit{“emotional”}). I calculated the sum of the responses to the six traits to form the \textit{GSE index}. The internal consistency across questions remained identical after the subselection of the questions ($\alpha = .88$). The GSE index had a possible range of [6, 54], in which a higher score indicates higher level of stereotype endorsement. The analyses with the total GSE score, which included all items, not just 6 items used for the GSE index, yielded consistent results (see Appendix E for details).

\textbf{Study 1b: Results and Discussion}

Before reporting the main results, I report how the present data replicates the findings of Oosterhof and Todorov (2008; Study 1a). In Study1b, when the impression ratings were collapsed across face genders, PC1 and PC2 accounted for over 80\% of the variance in the ratings newly collected for Study 1b (65.57\% and 19.90\%, respectively). PC1 and PC2 were loaded highly on by trustworthiness (.95) and dominance ratings (.94), respectively. This replicates Oosterhof and Todorov (2008), where PC1 and PC2 explained >80\% of the variance in the ratings (63.3\% and 18.3\%, respectively) and were loaded highly on by trustworthiness (.92) and dominance ratings (.87). When the impression ratings were separated depending on the face gender, again, the PCA solution of each gender from the present data replicated those in Oosterhof and Todorov (2008; Study 1a). Specifically, I ran correlational analyses of the PCA loadings of the social traits 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 \((R = .98)\). These findings suggest high stability of the impressions of male and female faces over time, since the ratings in Oosterhof and Todorov (2008) were collected over ten years ago.

As in Study 1a, for both genders, PC1 was highly loaded on by all positive and negative traits (Appendix B). This is consistent with previous models of face impressions (Oosterhof & Todorov, 2008; Sutherland et al., 2013) and Study 1a. Again, female face impressions were less differentiated and more valence-laden than male face impressions. As in Study 1a, the absolute value of the correlations was bigger for female than for male face ratings \((t(90) = 2.16, P < .05; \text{ Fig. 1})\), indicating a higher level of dependency between impressions in female face impressions. This is consistent with a visual inspection of the PCA solutions: the trait loadings on PC1 have bigger absolute values for female than for male faces (Appendix B). Correspondingly, PC1 explained a larger amount of variance in female face ratings than in male face ratings \((67.94\% \text{ vs. } 61.66\%; \text{ Fig. 2})\). These findings replicate the findings of Study 1a.

**Role of Raters’ Gender Stereotypes.** To examine how the rater GSE was related to gender differences in impression differentiation, I repeated the two main analyses (i.e., the correlational analyses and PCAs) for each face gender using responses of multiple subsets of raters, which I varied according to their GSE indices. Specifically, starting from the raters with the lowest GSE index (i.e., raters who endorsed gender stereotypes the least), I sub-selected raters such that there were always \(\geq 9\) raters for each impression rating in each participant subset (i.e., at least 126 raters = 9 raters \(\times 14\) impression), then I
varied the level of GSE of the sub-selected raters. As before, the average rating per face was calculated for each trait (for each subset of raters). I then conducted (a) a paired $t$-test on the absolute values of the inter-impression correlational coefficients between face genders and (b) a PCA on impression ratings of each face gender. Additionally, to understand the relationship between the impression differentiation and the rater GSE, I ran linear and quadratic regressions predicting the absolute pairwise correlational coefficients of male and female ratings using the raters’ GSE index as a predictor.

I measured the level of raters’ GSE by asking them about the extent to which they agree with the differences between men and women on gender-stereotypical traits (Appendix C). Each rater’s GSE index was calculated as the sum of their responses to a subset of questions with the highest inter-item reliability. The overall GSE index was higher than the absolute middle score (i.e., 30 in the range [6,54]) of the questionnaire ($M = 39.63$, $SD = 6.71$; $t(451) = 30.50$, $P < .001$), indicating that participants on average endorsed gender stereotypical beliefs (i.e., “men are more likely to be competitive/dominant/aggressive than women, and women are more likely to be nurturing/sensitive/emotional than men”). Male raters showed a higher level of gender stereotype endorsement ($M = 40.92$, $SD = 7.23$) than female raters ($M = 38.37$, $SD = 5.91$; $t(428.52) = 4.09$, $P < .001$).

Overall, as the rater GSE index increased, the correlations between impression ratings increased too for both male and female faces (Fig. 3), but the relationship was nonlinear. Although the linear regression model was significant for the ratings of both genders (male faces: $R^2 = .69$, $F(1,14) = 30.49$, $P < .001$; female faces: $R^2 = .59$, $F(1,14) = 19.96$, $P < .001$), so was the quadratic model (male faces: $R^2 = .98$, $F(2,13) = 269.5$, $P$
< .001; female faces: $R^2 = .98, F(2,13) = 415.8, P < .001$). The difference between the absolute correlation coefficients for male and female face ratings was significant in the subset of raters within the GSE index range of [35,39], as shown in the grey box in Figure 3A. That is, raters who more strongly endorsed gender stereotypes were more likely to show less differentiated impressions of female faces than of male faces, but this effect was not present for the raters with the strongest gender stereotypes. Correspondingly, the amount of variance explained by PC1 in the ratings followed the same quadratic pattern of change across the GSE index (Fig. 3B).

---

**Fig. 3.** The level of intercorrelations among impressions (A) and the amount of variance in the impressions explained by PC1 (B) as a function of the raters’ gender stereotype endorsement (GSE) level. Each data point was calculated from a subset of raters with sliding windows on the level of rater GSE ($n_p \geq 126$ for each subset), in which the x value is the middle point of the sliding window. The shades surrounding data points denote ±SE, and the rectangular shades denote a significant difference between the intercorrelations for male and female faces ($P < .05$). The semitransparent curves show quadratic fits of the data from the ratings of male (lighter grey) and female faces (darker grey).
Role of Raters’ Gender. To examine how the rater gender was related to differences in impression differentiation of male and female faces, I calculated correlational coefficients across impression ratings separately for male and female raters. I then conducted a 2 [face gender] × 2 [rater gender] repeated measures ANOVA on the absolute values of the inter-impression correlational coefficients. Additionally, I conducted PCAs separately for male and female faces, this time, using the mean impression ratings of male raters and using the mean ratings of female raters.

For both male and female raters, PC1 explained more variance in female than in male face impressions (55.72% vs. 54.18% in male raters, 65.69% vs. 60.77% in female raters). The 2 [face gender] × 2 [rater gender] repeated measures ANOVA on the absolute values of the intercorrelational coefficients between traits 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 ($M = 0.59, SD = 0.23$) than the male raters ($M = 0.50, SD = 0.23$), 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 ($M = 0.61, SD = 0.21$) vs. male faces ($M = 0.56, SD = 0.25; t(90) = 2.50, P < .025$, Bonferroni correction), whereas male raters showed the same level of cross-trait correlations for female ($M = 0.51, SD = 0.21$) vs. male faces ($M = 0.50, SD = 0.25; t(90) = 0.41$). This finding suggests that the less differentiated impressions of female faces are primarily due to female raters.
Study 1c: Introduction

Studies 1a and 1b both showed that impressions of women are less differentiated and more highly valence-laden compared to impressions of men. However, Studies 1a and 1b used the same face image set. To test the robustness of the results of the previous studies, in Study 1c, I ran the same analyses (i.e., correlational analyses, PCAs) on a preexisting face rating dataset involving different sets of face images, impressions, and participants (Ma, Correll, & Wittenbrink, 2015).

Study 1c: Methods

Participants. For the impression trait ratings previously collected by Ma et al. (2015), +1,087 participants had rated face images on various social traits.

Stimuli. Naturalistic face photos with direct gaze and neutral expressions were used. The individuals were amateur actors with no facial hair, earrings, eyeglasses, or visible make-up, all wearing grey T-shirts: 597 photos of 109 Asian (57 females), 197 Black (104 females), 108 Latino (56 females) or 183 White actors (90 females) between the ages of 17–65 were used (Ma et al., 2015).

Procedure. Participants rated 290 male and 307 female face photos on 16 social traits: how afraid, angry, attractive, babyfaced, disgusted, dominant, feminine, happy, masculine, racially prototypical, sad, surprised, threatening, trustworthy, or unusual each individual looked. To collect the face impression ratings, Ma et al. (2015) asked each participant to form impressions of individuals from photos on multiple attributes (e.g., “Consider the person pictured above and rate him/her with respect to other people of the same race and gender. - Fearful/Afraid (1=Not at all; 7=Extremely)”). Their dataset
consisted of multiple attribute ratings per face, each of which was averaged across raters. As in Studies 1a and 1b, I conducted two main analyses to test the gender differences in face impressions. First, to compare the in/dependency of perceived traits between genders, I ran a paired $t$-test between the absolute values of the coefficients for male and female face ratings. Correlational coefficients between the ratings of every possible trait pair among all 15 traits $(C(15,2) = 105$ pairs$)$ were calculated on the face image level for each gender. Second, to test whether raters showed less differentiated impressions for female than for male faces, I compared the amount of the variance explained by the first component for each gender.

**Study 1c: Results and Discussion**

As in Studies 1a and 1b and previous models of face impressions, for both genders, PC1 was highly loaded on by all positive and negative traits (Appendix B). Again, female face impressions were less differentiated and more valence-laden than male face impressions. As in Studies 1a and 1b, the absolute value of the correlations was bigger for female than for male face ratings ($t(104) = 3.30, P < .01; \text{Fig. 1}$), indicating a higher level of dependency between traits in female face impressions. This is consistent with a visual inspection of the PCA solutions: the trait loadings on PC1 have bigger absolute values for female faces than for male faces (Appendix B). Correspondingly, PC1 explained a larger amount of variance in the female than in male ratings (40.87% vs. 31.60%; Fig. 2), indicating a higher level of dependency of female ratings on valence. Overall, the results replicate what I found in Studies 1a and 1b: face impressions of women are less differentiated than those of men, and are highly valence-laden.
Part I: Interim Conclusion

In Part I, I found that people hold less differentiated (i.e., more highly intercorrelated) trait impressions of women than of men. Specifically, impressions of women varied from each other to a smaller degree and were more tied to overall positivity/negativity evaluation than impressions of men (Studies 1a – 1c). The current findings confirm ambivalent sexism (Glick & Fiske, 1996; 2011) in the domain of visual perception. The theory of ambivalent sexism posits that women are evaluated positively as far as they are stereotyped into restricted traits and roles (e.g., being helpful), unlike men whose impression valence is less dependent on stereotypes. This theory would posit that women with counterstereotypical looks are more likely to perceived as less typical and evaluated more negatively than men with counterstereotypical looks. This gender difference will lead to a higher level of intercorrelations between the impressions of women than between those of men. This is consistent with present findings. In the same vein, the current findings suggest that the backlash effect, a phenomenon in which women who violate prescriptions of feminine traits receive social or economic penalties (Rudman, 1998; Rudman & Glick, 2001; Rudman & Phelan, 2008), generalizes to impressions based on facial appearance, corroborating previous research (Sutherland et al., 2015). Specifically, these results corroborate previous research showing that qualities perceived as traditionally masculine led to more negative valence-related impressions of women (e.g., likability). In these studies, women with dominant facial looks (Sutherland et al., 2015), assertive attitude (Rudman, 1998; Rudman & Glick, 2001; Rudman & Phelan, 2008), or work competency (Hagen & Kahn, 1975) were evaluated more negatively than women with the opposite qualities. Although qualities perceived as
traditionally feminine sometimes lead to more negative valence-related impressions of men (Derlega & Chaikin, 1976; Heilman & Wallen, 2010; Moss-Racusin, Phelan, & Rudman, 2010), the effects of counterstereotypical looks on impressions of men are weaker (Sutherland et al., 2015) and sometimes even beneficial. For instance, for both men and women, a feminine face shape is perceived as attractive (Little, Burt, Penton-Voak, & Perrett, 2001; Penton-Voak et al., 1999; Perrett et al., 1998; Rhodes, Hickford, & Jeffery, 2000; Said & Todorov, 2011; but see Rhodes, 2006). In sum, when gender prescriptive norms (i.e., behavior/traits considered desirable for a gender) are violated, women face harsher penalties than men on average. Moreover, traits that are considered desirable or typical for women outnumber the traits considered desirable or typical for men (Prentice & Carranza, 2002). These asymmetries seem to lie at the core of the less differentiated face-based impressions of women.

The level of impression differentiation was affected by the characteristics of those who formed the impressions. Raters who were more likely to endorse gender stereotypes showed less differentiated impressions of both men and women than raters who were less likely to endorse these stereotypes (Study 1b). This finding supports the idea that gender differences in impression differentiation result from evaluative processes triggered by gender categorization and, more generally, the idea that trait impressions of a group are strongly or weakly intercorrelated depending on the perceiver’s stereotypes of the group (Secord & Berscheid, 1963; Stolier, Hehman, & Freeman, 2017). For those who endorse gender stereotypes, gender-consistent appearance would lead to more positive impressions, whereas gender-inconsistent appearance would lead to more negative impressions. These processes could lead to stronger intercorrelations between
impressions in those who strongly endorse gender stereotypes. These individuals may evaluate others based on their gender category because their gender stereotypes are highly accessible (Higgins, 1996). Given this inference, it is puzzling that raters with the strongest GSE did not show the strongest intercorrelations between impressions, as demonstrated by the quadratic relationship between the rater GSE and the impression differentiation. More research is needed to determine the underlying causes.

Participant gender also affected the level of impression differentiation. It was female raters who showed less differentiated impressions of women. This participant gender effect may seem surprising because female participants are less likely to endorse gender stereotypes or traditional gender roles compared to male participants (Glick & Fiske, 1996; Swim, Aikin, Hall, & Hunter, 1995; Williams & Best, 1990b). However, female participants sometimes do show evaluation consistent with gender stereotypes, such as negative evaluation of other females with counterstereotypical traits (Garcia-Retamero & López-Zafra, 2006; Goldberg, 1968; Parks-Stamm, Heilman, & Hearns, 2007; Rudman, 1998). This may explain the lower level of impression differentiation of female participants for female than for male faces.

All in all, in Part I, across multiple datasets, I found a low level of impressions differentiation for female faces, which was predicted by perceivers’ gender stereotypes. Where does the face-gender difference in impressions stem from? To this question, in the next part, I model the impressions of male and female face impressions.
Part II: Gender Biases in Models of Impressions

In Part I, I found that face impressions of women are less differentiated and are more highly valence-laden than those of men. These findings suggest that impressions need to be understood separately for each gender rather than being collapsed across gender as in previous work (e.g., Oosterhof & Todorov, 2008). In Part II, I built and validated models of male and female impressions to understand the perceptual basis of the gender differences in impressions observed in Part I. Using these models, in other words, I studied to what extent the gender difference in impressions is related to how people use facial information when forming impressions (e.g., using facial masculinity to form an impression of dominance). Specifically, people may either (a) use the same facial information for both genders to form an impression but interpret the information differently or (b) use different facial information for different genders.

To investigate the extent to which people use dis/similar facial information to form impressions of different genders, in Part II, I built data-driven computational models of impressions separately for men’s and women’s faces, and calculated the similarities between these gender-specific face impression models. Each model represents what facial information is used to form a single impression for a particular face gender (e.g., an impression of dominance of male faces). One way to assess the similarities in the facial information used for impressions between face genders is to directly study the correlational similarities across face impression models (Study 2). Another way is to test how well these face impressions models can manipulate impressions of male and female faces within and across genders (Study 3). In the next section, I explain what exactly these analyses test and how these analyses were conducted in concrete terms.
Study 2: Introduction

To investigate the extent to which people use dis/similar facial information to form impressions of men and women, in Study 2, I built data-driven models of impressions separately for male and female faces. I then calculated the similarities between the gender-specific face impression models using correlational analyses between the models. Data-driven face modeling reveals facial features that correlate with an impression with little prior assumptions of what features matter (Funk, Walker, & Todorov, 2016; Jack & Schyns, 2017; Oosterhof & Todorov, 2008; Todorov et al., 2011; Todorov & Oosterhof, 2011; Walker & Vetter, 2009; 2016). In prior work, the participants rated a trait impression of randomly generated faces (e.g., how dominant they looked) from a multidimensional, statistical face space, where each face is represented as a point in this space (Oosterhof & Todorov, 2008). In this approach, each impression model is a vector in the face space visualizes holistic visual changes on a face as it looks more trait-like (e.g., more dominant). All prior models were built irrespective of the gender of faces. Here, I built separate gender models of impressions of trustworthiness and dominance.

Study 2: Methods

Participants. Five-hundred-and-ten online participants (233 males, 256 females, 21 unreported) participated through Amazon MTurk for monetary reward.

Stimuli. Previous data-driven computational models of face impressions were based on impression ratings of randomly generated faces from a multidimensional, statistical face space (Oosterhof & Todorov, 2008). I built new models separately for male and female
faces. I generated 301 male faces and 301 female faces with FaceGen 3.2 (Singular Inversions). FaceGen model is based on a database of male and female human faces that were laser-scanned in 3D. In the model, a face is represented as a point in a multi-dimensional face space. Moving a point (i.e., a face) along a dimension results in a holistic change in the shape or reflectance (i.e., texture and coloration) of the face in a specific way. The shape and the reflectance of a face are determined by 50 shape and 50 reflectance parameters on 100 dimensions, respectively. I generated male and female faces (300 each) by randomly selecting each parameter from a normal distribution. I used a single set of 300 source faces and made them more male- or female-like, so that all the male and all the female faces were centered around the average male and the average female faces in FaceGen database, respectively. The male and female average faces were also included in the stimulus set. This resulted in paired images of male and female faces (see Appendix G for sample faces).

Procedure. To build gender-specific models of impressions, I asked participants to rate a random set of computer-generated male or female faces on trustworthiness and dominance. I chose these two impressions because they are the best approximations of the valence and power dimensions underlying face impressions and they are highly distinctive from each other (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov, Said, Engell, & Oosterhof, 2008). Then, I built statistical models of the two trait impressions based on the participants’ ratings.

Each participant rated either 51 male or 51 female faces on one of two impressions: trustworthiness or dominance. The 51 faces consisted of 50 random faces from the pool of 300 random faces and the gender-specific average face. Participants
were told that there were no right or wrong answers and that the research was interested in their first impression or “gut response”. The faces were presented twice in two separate blocks for the reduction of the measurement error and the screening of unreliable raters’ data. I calculated each rater’s intra-rater reliability by correlating their ratings from different blocks. The ratings from participants with zero or negative intra-rater reliability were excluded, which resulted in responses of 418 participants. Each face image was presented with a question (“How [trait term, e.g., trustworthy] is this [man/woman]?”) and a response scale below the face (“1 Not at all [trait term] - 9 Extremely [trait term]”). Each face was visible until the participant responded, the ITI was 250 ms, and the order of faces within a block was randomized. The ratings between trustworthiness and dominance were negatively correlated for both genders, but the correlation was stronger for female faces (see Appendix H for a scatterplot of the ratings).

To create gender-specific computational trait models (i.e., male trustworthiness model, female trustworthiness model, male dominance model, and female dominance model), I averaged the ratings per face across raters for each trait. For each gender-specific model, I computed the contribution of each of the 50 shape and the 50 reflectance parameters to the average trait impression ratings of the 300 faces, following the previous data-driven statistical approach (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). The mean impression ratings of the 301 faces and the values for a single parameter (out of 100) of the 301 faces are essentially two vectors with 301 elements each. To create one parameter of an impression model, the cross-product of these two vectors were summed across faces, and then were normalized across
parameters. The resulting 100 parameters of the model represent the amount of variation that would induce a 1SD change in the impression rating.

**Study 2: Results and Discussion**

The resulting gender-specific statistical impression models are shown in Figures 4 – 7. Both models derived from male and female face ratings are similar to existing statistical models of impressions (Oosterhof & Todorov, 2008). This finding, again, as Study 1b, suggests high stability of facial impressions over time, since the impression models in Oosterhof and Todorov (2008) were derived from responses collected over ten years ago. Further, the gender-specific models represent similar facial information found in prior research: as both male and female faces are manipulated to appear more trustworthy, their expressions become more positive and vice versa (Figs. 4 – 5; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009). Likewise, as both male and female faces are manipulated to appear more dominant, they become more masculine and facially mature (Figs. 6 – 7; Oosterhof & Todorov, 2008; Sutherland et al. 2013; Zebrowitz, 2005; Zebrowitz & Montepare, 2008).

Within each impression, male and female models were highly positively correlated, suggesting that similar information is used when people form impressions of male and female faces (trustworthiness: $\rho = .68$, dominance: $\rho = .85$; see Appendix I for all indices of similarities). However, trustworthiness and dominance models were more strongly negatively correlated in the female ($\rho = -.38$) than in the male models ($\rho = -.16$), suggesting that these models are more similar for female than for male faces. This
confirms the empirical data reported in Studies 1 and 2, in which face impressions of women were less differentiated than those of men.

**Role of Raters’ Gender.** The correlation between female trustworthiness and dominance models were stronger for female ($\rho = -0.42$) than male raters ($\rho = -0.30$), whereas the correlation between male trustworthiness and dominance models were comparable for female ($\rho = -0.16$) and male raters ($\rho = -0.16$). This finding shows that impressions of trustworthiness and dominance of female faces relied on more similar visual information for female than for male raters. This is consistent with the empirical data reported in Study 1b, in which female raters showed less differentiated impressions for female than for male faces.
Fig. 4. Models of female (top) and male (bottom) trustworthiness impression applied to a sample male face.

Fig. 5. Models of female (top) and male (bottom) trustworthiness impression applied to a sample female face.
Fig. 6. Models of female (top) and male (bottom) dominance impression applied to a sample male face.

Fig. 7. Models of female (top) and male (bottom) dominance impression applied to a sample female face.
Study 3a: Introduction

Study 2 showed that people use similar facial information to form an impression of men and women, as reflected in high level of correlation between genders (i.e., trustworthiness: $\rho = .68$, dominance: $\rho = .85$). To formally test whether people use the same information to form an impression of men and women, in Study 3, I cross-validated the gender-specific impression models using novel male and female faces (i.e., test whether the models explain people’s impressions). I manipulated the level of perceived traits of novel images of male and female faces (see Appendix K for the novel original face images), and asked participants to rate the faces on the respective traits. The first objective of this study was to test whether the models of impressions successfully capture the changes in facial appearance that lead to changes in impressions. The second and more important object was to test whether the gender specific models work better when applied to a congruent face (e.g., a male model applied to a male face). The latter possibility, despite the similarity of the gender-specific models observed in Study 2, would suggest that people use different visual information when forming impressions of men and women. The alternative possibility is that all models, irrespective of whether they are derived from ratings of male or female faces, would be equally capable of manipulating impressions of both male and female faces. Such a result would suggest that people use very similar information when forming impressions of male and female faces although they interpret this information differently (e.g., whereas masculine features on a male face are evaluated positively, masculine features on a female face are evaluated negatively).
Study 3a: Methods

Participants. Two-hundred-and-sixty-eight online participants (121 males, 147 females) participated through Amazon MTurk for monetary reward.

Stimuli and Procedure. To validate the gender-specific face models, I generated faces that reflected the impression change in each model. First, using a procedure similar to previous validation studies (Todorov et al., 2013), I generated 25 new faces. To obtain these faces, I generated 1,000 faces whose positions on the 100 parameters were independently sampled from 100 normal distributions. From the 1,000 random faces, 25 faces that physically differed maximally to one another were chosen (i.e., faces with highest average Euclidean distance to each other; Appendix K). I used a single set of 25 faces and made them more male- or female-like.

Second, I manipulated each face with the social trait models. I varied the face parameters of the 50 faces by adding -3, -2, -1, 0, 1, 2, and 3SDs on each model, with the 0SD addition being null manipulation. There were 4 gender-specific social trait models, i.e., male trustworthiness model, female trustworthiness model, male dominance model, and female dominance (Figs. 4 – 7). This resulted in 1,400 faces (2 [face gender] * 25 [identities] * 7 [manipulation levels] * 4 [model]).

Each participant rated either male or female faces manipulated by either male or female model of either trustworthiness or dominance. Participants were told that there were no right or wrong answers and that the research was interested in their first impression. For each participant, 175 faces were presented first (25 [identities] * 7 [levels]), followed by presentation of 25 randomly chosen faces from the previously presented faces without any noticeable break. The 25 faces were repeated for the
calculation of test-retest reliability. The ratings from participants with zero or negative intra-rater reliability were excluded, which resulted in responses of 247 participants. Each face gender $\times$ model trait $\times$ model gender condition had $\geq$ 30 raters. The ratings of trustworthiness and dominance were highly reliable irrespective of whether the gender of the original faces and the model were identical ($\alpha_{\text{min}} = .96$) or not ($\alpha_{\text{min}} = .96$; see Appendix J for all $\alpha$’s).

**Study 3a: Results and Discussion**

To assess how well the models varied the intended impressions of faces, I ran linear and quadratic regressions for the trait ratings of male and female faces with the level of model manipulation as the predictor. Then, to determine whether the gender-specific models are more successful when applied to a congruent face (e.g., a female model applied to a female face), I compared the predictive powers of the models across genders. I ran $2 \times 2 \times 2$ repeated measures ANOVAs on Fisher’s $z$ scores, which were transformed from the correlations between the observed ratings and the predicted ratings of the linear and quadratic regressions.

All linear and quadratic models explained significant amount of variance in the impression ratings regardless of whether the model gender and the face gender were congruent (linear: $R^2$s $> .94$, quadratic: $R^2$s $> .96$; Fig. 8) or not (linear: $R^2$s $> .94$, quadratic: $R^2$s $> .97$). Thus, the effectiveness of the trait model was not affected by the congruency between the face gender and the model gender.
Fig. 8. Validation of models of trustworthiness (top) and dominance (bottom) with computer-generated male (left) and female faces (right). Linear (gray) and quadratic (black) fit of ratings of trustworthiness as a function of the female (solid) and male (dashed) model values of the faces. Overall, models generated faces varying on the intended impressions within and across gender, suggesting little gender specificity in facial information used to form each impression. Error bars denote ±SE.

To further assess the relative effectiveness of the models, I ran a 2 [face gender] × 2 [model trait] × 2 [model gender] repeated measures ANOVA on the correlations between the predicted and observed ratings (for the analysis, the correlations were transformed to Fisher’s z scores) for both the linear and quadratic regression models. I found a main effect of model trait for the linear ($F(1,24) = 5.19, P < .05, \eta^2_G = .05$) and
the quadratic regression models \(F(1,24) = 5.98, P < .05, \eta_G^2 = .03\). Specifically, the dominance models predicted the ratings better (linear: \(M = 2.42, SD = 0.36\); quadratic: \(M = 2.75, SD = 0.42\)) than the trustworthiness models did (linear: \(M = 2.26, SD = 0.35\); quadratic: \(M = 2.63, SD = 0.35\)). I also found a main effect of face gender for the linear \(F(1,24) = 10.52, P < .01, \eta_G^2 = .04\) and quadratic regression models \((F(1,24) = 4.40, P < .05, \eta_G^2 = .02)\). Specifically, the ratings of male faces were better predicted by the models (linear: \(M = 2.41, SD = 0.37\); quadratic: \(M = 2.74, SD = 0.40\)) than the ratings of female faces (linear: \(M = 2.28, SD = 0.35\); quadratic: \(M = 2.65, SD = 0.38\)). These two effects were not predicted and, as shown in Figure 8, they were relatively small.

Only for the quadratic regression model, I found the significant model gender \(\times\) face gender interaction \((F(1,24) = 7.09, P < .05, \eta_G^2 = .03)\), indicating that the male models were better at manipulating impressions of female faces \((M = 2.75, SD = 0.42)\) than male faces \((M = 2.71, SD = 0.32)\), whereas the female models were better at manipulating impressions of male faces \((M = 2.76, SD = 0.47)\) than female faces \((M = 2.54, SD = 0.32)\). In sum, the models could generate faces varying on the intended impressions within and across gender. The effect of gender congruency was only found in the quadratic models, the direction of the effect was the opposite of what was hypothesized, and the effect size was relatively small.

These findings suggest that people use highly similar information between face genders. To assess the robustness of the results, in Study 3b, I conduct another validation study using real-life face images, not computer-generated face images.
Study 3b: Introduction

In Study 3a, I found no evidence for gender specificity in the models’ capacity to manipulate impression. If anything, for the quadratic model, the predicted interaction was in the opposite direction. These findings suggest that people may be using the same visual information when forming facial impressions of male and female faces. However, Study 3a used computer-generated synthetic faces to test the gender specificity of the facial information in impression formation. People might form first impressions differently when presented with real-life faces. Real-life face images are different from synthetic face images in significant ways. For example, real-life faces are less ambiguous in gender than synthetic faces. Thus, it may be inadequate to conclude that people use highly similar information to form an impression of male and female faces only based on the validation results with synthetic face images. In Study 3b, by using photorealistic face stimuli, I test whether the null effect of the face gender reported in Study 3a generalizes to real-life face images. Specifically, I manipulated the level of perceived traits of novel real-life images of male and female faces with gender-specific trait models. I then asked participants to rate the faces on the respective traits.

Study 3b: Methods

Participants. Two-hundred-and-ninety-two online participants (127 males, 164 females, 1 other gender) participated through Amazon MTurk for monetary reward.

Stimuli and Procedure. Photos of male and female faces (25 each) were randomly selected from a face database (DeBruine & Jones, 2017), consisting of naturalistic face photos with direct gaze and neutral expressions without any eyeglasses or visible make-
up, all wearing white T-shirts. In total, 8 photos of East Asian (4 females), 8 photos of West Asian (3 females), 12 photos of Black (5 females), and 22 photos of White actors (13 females) between the ages of 19 – 37 were used.

Using the impression models, I manipulated each face along the respective trait impression. As in Study 3a, I prepared seven facial variations, including the original face, for each face identity × impression model. To apply the models, I transformed the initial face images along the gender-specific impression models from Study 2, using PsychoMorph (Tiddeman, Burt, & Perrett, 2001). First, I created synthetic faces that represented these models, extreme faces that are -4 and 4SD deviant from the average FaceGen face on each model. Next, I used the transformation function of PsychoMorph to change the initial 25 male and 25 female real-life face images along the continuum of the difference between the two extreme face images, for each impression. Unlike the standard morphing procedure, which is a direct transition between two images, the transformation procedure in PsychoMorph allows users to manipulate a single image along a continuum and generate photo-realistic images (Sutherland, Rhodes, & Young, 2017). On the most extreme ends, each face image was transformed 40% away from the original face. The value of 40% was chosen because any stronger manipulation than this value caused a distortional artifact on face images. The manipulation magnitude was identical for the intervals between the 7 facial variations: the final face images were transformed -40.00%, -26.67%, -13.33%, 0%, 13.33%, 26.67%, and 40.00% towards the extreme faces from the initial faces (Figs. 9 – 12). To maintain the gender and ethnicities of the original faces, I only used the variation in the face shape of the models. As in
Study 3a, there were 4 gender-specific social trait models. This resulted in 1,400 faces (2 [face gender] × 25 [identities] × 7 [manipulation levels] × 4 [model]).

As in Study 3a, each participant rated either male or female faces manipulated by either male or female model of either trustworthiness or dominance. Participants were told that there were no right or wrong answers and that the research was interested in their first impression. For each participant, 175 faces were presented first (25 [identities] × 7 [levels]), followed by presentation of 25 randomly chosen faces from the previously presented faces without any noticeable break. The 25 faces were repeated for the calculation of test-retest reliability. The ratings from participants with zero or negative intra-rater reliability were excluded, which resulted in responses of 239 raters. Each face gender × model trait × model gender condition had ≥30 raters. The ratings of both trustworthiness and dominance were highly reliable irrespective of whether the gender of the original faces and the model were identical ($\alpha_{\text{min}} = .88$) or not ($\alpha_{\text{min}} = .83$; see Appendix L for all $\alpha$’s).

**Study 3b: Results and Discussion**

I adopted the same procedure and analyses as those in Study 3a: regressions for the trait ratings and 2 [face gender] × 2 [model trait] × 2 [model gender] repeated measures ANOVAs on Fisher’s $z$ scores converted from the correlations between the observed ratings and the predicted ratings of the regressions.
Fig. 9. Models of female (top) and male (bottom) trustworthiness impression applied to a sample male face. Each number represents the extent to which the original face was transformed towards one of the two extreme faces that represent the impression model.

Fig. 10. Models of female (top) and male (bottom) trustworthiness impression applied to a sample female face. Each number represents the extent to which the original face was transformed towards one of the two extreme faces that represent the impression model.
Fig. 11. Models of female (top) and male (bottom) dominance impression applied to a sample male face. Each number represents the extent to which the original face was transformed towards one of the two extreme faces that represent the impression model.

Fig. 12. Models of female (top) and male (bottom) dominance impression applied to a sample female face. Each number represents the extent to which the original face was transformed towards one of the two extreme faces that represent the impression model.
Fig. 13. Validation of models of trustworthiness (top) and dominance (bottom) with real-life male (left) and female face images (right). Linear (gray) and quadratic (black) fit of ratings of trustworthiness as a function of the female (solid) and male (dashed) model values of the faces. Overall, models generated faces varying on the intended impressions within and across gender, suggesting little gender specificity in facial information used to form each impression. Error bars denote ±SE.

As in Study 3a, all models explained significant amount of the variance in the impression ratings regardless of whether the model gender and the face gender were congruent (linear: $R^2$s > .54, quadratic: $R^2$s > .68; Fig. 13) or not (linear: $R^2$s > .63, quadratic: $R^2$s > .68). To further assess the relative effectiveness of the models, I ran a 2 [face gender] × 2 [model trait] × 2 [model gender] repeated measures ANOVA on Fisher’s $z$ scores converted from the correlations between the predicted and observed
ratings for both the linear and quadratic regression models. I found a significant main effect of model trait for the linear \((F(1,24) = 82.73, P < .001, \eta^2_G = .35)\) and the quadratic regression models \((F(1,24) = 75.49, P < .001, \eta^2_G = .33)\). Specifically, the dominance models predicted the ratings better \((\text{linear}: M = 1.87, SD = 0.45; \text{quadratic}: M = 2.11, SD = 0.48)\) than the trustworthiness models did \((\text{linear}: M = 1.22, SD = 0.49; \text{quadratic}: M = 1.46, SD = 0.49)\). The same effect was observed in Study 3a.

I also found a main effect of model gender for the linear \((F(1,24) = 4.67, P < .05, \eta^2_G = .02)\) and the quadratic regression models \((F(1,24) = 4.93, P < .05, \eta^2_G = .02)\). Specifically, male faces were more effectively manipulated by impression models \((\text{linear}: M = 1.61, SD = 0.52; \text{quadratic}: M = 1.86, SD = 0.51)\) than female faces \((\text{linear}: M = 1.49, SD = 0.61; \text{quadratic}: M = 1.71, SD = 0.64)\). I found the significant model trait \(\times\) face gender interaction for the linear \((F(1,24) = 7.84, P < .01, \eta^2_G = .03)\) and the quadratic regression models \((F(1,24) = 8.66, P < .01, \eta^2_G = .03)\), indicating that dominance models were more effective at manipulating female \((\text{linear}: M = 1.91, SD = 0.50; \text{quadratic}: M = 2.22, SD = 0.52)\) than male faces \((\text{linear}: M = 1.84, SD = 0.39; \text{quadratic}: M = 2.00, SD = 0.41)\), whereas the trustworthiness models were more effective at manipulating male \((\text{linear}: M = 1.34, SD = 0.49; \text{quadratic}: M = 1.51, SD = 0.53)\) than female faces \((\text{linear}: M = 1.09, SD = 0.45; \text{quadratic}: M = 1.42, SD = 0.46)\). None of these effects were observed in Study 3a.

Across the two validation analyses with computer-generated and real-life faces, the only reliable finding was that the dominance impression models could generate faces varying on the intended impressions better than the trustworthiness impression models could, irrespective of the face gender or the model gender. All in all, the models were all
capable of generating faces varying on the intended impressions within and across gender, showing no evidence for gender specificity.

**Part II: Interim Conclusion**

In Part II, I found that the models of trustworthiness and dominance impressions of men and women were based on similar facial information. These are the first computational models of trustworthiness and dominance impressions built separately for men and women. Correlational analyses (Study 2) and cross-gender validation of these models (Study 3) consistently showed that people use similar visual information to form impressions of both genders. Combined with the previous finding that first impressions from faces are less differentiated for women than for men (Part I), this finding suggests that people use the same visual information when forming impressions of men and women, but that this information is evaluated differently, emphasizing the role of the face gender in impression formation. Specifically, what kind of facial information is used to form impressions of both genders?

Inspecting the visualizations of the models (Figs. 4 – 7 and 9 – 11) showed that resemblance to emotional expressions is a key input to trustworthiness impressions (Adams, Nelson, Soto, Hess, & Kleck, 2012; Engell, Todorov, & Haxby, 2010; Hess, Blairy, & Kleck, 2000; Keating, Mazur, & Segall, 1981; Montepare & Dobish, 2003; Oosterhof & Todorov, 2008; 2009; Said, Sebe, & Todorov, 2009; Sutherland et al., 2013; Zebrowitz & Montepare, 2008) for both male- and female-based models. Consistent with prior models (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Walker & Vetter, 2009), as the faces are manipulated to look more trustworthy, they acquire more positive expressions. In contrast, as the faces are manipulated to look less trustworthy, they
acquire more negative expressions. For dominance impressions, the key inputs are masculinity and facial maturity (Oosterhof & Todorov, 2008; Sutherland et al., 2013; Zebrowitz, 2005; Zebrowitz & Montepare, 2008) and, to a smaller extent, similarity to angry facial expressions (Adams et al., 2012; Hareli, Shomrat, & Hess, 2009; Hess et al., 2000; Hess, Adams, Grammer, & Kleck, 2009; Said et al., 2009). As the faces are manipulated to look more dominant, they become more masculine and mature, and they acquire more negative expressions. In contrast, as the faces are manipulated to look less dominant, they become more feminine and baby-like, and they acquire more positive expressions.

Comparing the male- and female-based models, it is also possible to see specific gender differences, especially for the models of trustworthiness impressions. For example, faces on the positive end of the female trustworthiness model are rounder and more light-skinned compared to faces on the positive end of the male trustworthiness model (Figs. 4 – 5), possibly reflecting real gender differences (Jablonski & Chaplin, 2000). However, as the cross-validation results showed, these gender differences did not seem to matter for impressions. Impressions of both male and female faces were successfully manipulated irrespective of whether the model was derived from ratings of female or male faces, and this was the case for synthetic and real-life faces (Study 3). These findings suggest that when forming impressions of men and women on trustworthiness and dominance, people use the same information, but they evaluate this information differently. Consistent with prior research (Freeman et al., 2011; 2015; Hugenberg & Bodenhausen, 2004; Kramer, Young, Day, & Burton, 2017; Stolier & Freeman, 2016; 2017; Sutherland et al., 2015),
the current findings highlight the importance of social categorization in person impressions.

The difference in the structure of impressions of men and women has implications for both social perception theories and social justice. Although often visually ambiguous and conceptually continuous, gender is thought of as a categorical variable (Fiske & Neuberg, 1990), and people show a high level of consensus regarding the gender typicality of faces (Hehman, Sutherland, Flake, & Slepian, 2017). Further, gender-related differences in facial features are easily detectable from faces (Burriss, Little, & Nelson, 2007; Schyns, Bonnar, & Gosselin, 2016), and the facial features that vary across genders are processed in the early stages of perception (Cellerino et al., 2007; Mouchetant Rostaing, Giard, Bentin, Aguera, & Pernier, 2000; Mouchetant-Rostaing & Giard, 2003; Welling, Bestelmeyer, Jones, DeBruine, & Allan, 2017). Moreover, facial features that correlate with gender (e.g., masculine features) shape the formation of person impressions (Mattarozzi, Todorov, Marzocchi, Vicari, & Russo, 2015; Oosterhof & Todorov, 2008; Sutherland et al., 2013; 2015). For social perception theories, the findings would suggest that gender categorization shapes person impression formation beyond associations between specific facial features and trait impressions; combining its influence with the influence of social status cues (Freeman, Penner, Saperstein, Scheutz, & Ambady, 2011) and cultural cues (Freeman et al., 2015). For social justice, the findings would suggest another contributing factor to gender discrimination. Women with counterstereotypical appearance are perceived more unfavorably and are discriminated against more harshly compared to men with counterstereotypical appearance (Rudman, 1998; Rudman & Phelan, 2008; Sutherland et al., 2015). Given the importance of first
impressions (Ballew & Todorov, 2007; Blair et al., 2004; Eberhardt et al., 2006; Funk & Todorov, 2013; Olivola & Todorov, 2010; Todorov et al., 2005), less differentiated impressions of women would result in evaluative inferences that may penalize women more strongly than men for not fitting the expected stereotypes.

In sum, people have less differentiated and more valence-laden impressions of women than of men, although these impressions are based on similar visual information. These findings of Part II, with those from Part I, suggest that discrimination against women can start from the moment of forming first impressions, as women with counterstereotypical looks are likely to be evaluated negatively.
Part III: Gender Biases in Impressions of Competence

So far, in Part I and Part II, I have observed that impressions of women are more simplified and heavily valence-laden compared to those of men (Part I) even though each impression is based on visual information that is similar between the two genders (Part II). What are some of the (potentially harmful) outcomes of gender biases in specific impressions, especially an impression with a significant real-world outcome? In Part III, I investigated competence impressions in relation to gender biases.

Impressions of competence are important, as they influence decisions about leadership selection (Antonakis & Eubanks, 2017). Intuitive judgments of competence from faces, for instance, can predict the results of political elections (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Lenz & Lawson, 2011; Little et al., 2007; Olivola & Todorov, 2010; Poutvaara, Jordahl, & Berggren, 2009; Todorov et al., 2005) and company executives’ compensation (J. R. Graham, Harvey, & Puri, 2017; Stoker, Garretsen, & Spreeuwers, 2016). It is important to understand the perceptual basis of these impressions because people act on these impressions (e.g., choose their leaders based on competence impressions) despite the dubious relationship between leaders’ actual competence and competence impressions from their faces (Stoker et al., 2016; Wyatt & Silvester, 2018).

Here, I investigated the visual ingredients of the competence stereotype. One of these ingredients is facial attractiveness. Both empirical studies and computational models of impressions support the halo effect of attractiveness (Dion, Berscheid, & Walster, 1972; Landy & Sigall, 1974; Thorndike, 1920) on competence impressions. First, a meta-analysis showed a modest to strong association between attractiveness and
perceived social/intellectual competence (Eagly, Ashmore, Makhijani, & Longo, 1991). Individuals with attractive faces are perceived as socially/occupationally competent (Dion et al., 1972; Landy & Sigall, 1974) and as having a higher social status (Webster & Driskell, 2015), which is strongly associated with perceived competence (Fiske et al., 2002). In real-world data, judgments of competence and attractiveness from politicians’ faces are highly correlated \((n = 244; \text{Olivola & Todorov, 2010})\). Second, data-driven models of impressions (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011) show a strong similarity between models of competence and attractiveness (Todorov et al., 2013). Since the two models exist in a common space, one can directly assess the similarities between them; indeed, the models of competence and attractiveness have been found to be highly similar \((\rho = .71)\), suggesting that people rely on attractiveness when forming impressions of competence.

In Part III, I tested whether any meaningful (potentially gender-biased) visual components other than attractiveness contribute to competence impressions. Data-driven computational models of impressions (Todorov et al., 2013; Walker & Vetter, 2016) are particularly suitable for addressing this question. Because the competence and attractiveness models are in the same statistical space, one can create a new competence model that is not confounded by attractiveness. One can either (a) make the new competence model to be uncorrelated with the attractiveness model by statistically making the model orthogonal to the attractiveness model, or (b) force the new model to be negatively correlated with the attractiveness model by subtracting the attractiveness from the competence model. To the extent that the new competence model (e.g., the resulting [competence – attractiveness] model) is meaningful, faces that are perceived as
more competent should not be perceived as more attractive. More importantly, if the model still predicts competence impressions, then by inspecting this model, one can find meaningful components of competence impressions that are not readily apparent. One potential component of these impressions is facial masculinity. When asked to evaluate the self and others on multiple attributes, people on average evaluate men as more competent (Bem, 1974; Rudman & Glick, 2012; Spence et al., 1975; Vogel et al., 1972) and confident (Spence et al., 1975; Vogel et al., 1972) compared to women. Further, the beliefs in the association between men and competence, confidence, and related traits (e.g., independence, inventiveness) are held across diverse cultures (Williams & Best, 1982; 1990a). However, the influence of masculinity on competence impressions may not be immediately apparent in the model of competence, because attractiveness is highly correlated with feminine facial appearance in both genders (Little et al., 2001; Penton-Voak et al., 1999; Perrett et al., 1998; Said & Todorov, 2011; but see Rhodes, 2006). By controlling for the attractiveness of faces, one can directly test whether masculinity contributes to competence impressions.

Following the abovementioned logic, I test for the gender biases in competence impressions and uncover multiple components underlying competence impressions. First, I showed that both judgments of attractiveness and competence change as faces were manipulated to look more competent by the standard competence model (Study 4a; Todorov et al., 2013). Then, I created a new model of competence by subtracting the effect of the attractiveness model from the original competence model and showed that faces that were manipulated by this model to look competent were indeed perceived to be more competent but not attractive (Study 4b). Next, I showed that these faces were also
perceived as more masculine and confident (Study 5). Further, I showed that these competent-looking faces were more likely to be categorized as men than as women and that the incompetent-looking faces were more likely to be categorized as women than as men (Study 6). Lastly, I extended the findings to real-life face images (Study 7). I showed that whereas masculinity cues increased competence impressions of male faces, they increased competence impressions of female faces only up to a point. Combined, this series of studies reveal how the way people form competence impressions are gender-biased.

**Study 4a: Introduction**

Using a validated, data-driven, computational model of competence (Todorov et al., 2013), I manipulated faces so that they varied in their perceived competence (Fig. 14, top). Participants were asked to evaluate these faces either on competence or on attractiveness. Given the high positive correlation between attractiveness and competence impressions (e.g., Dion et al., 1972; Todorov et al., 2013), I expected that as the level of model manipulation increases, both competence and attractiveness ratings of the faces would increase.

**Study 4a: Methods**

*Participants.* Thirty-three online participants (17 males, 16 females) participated through Amazon MTurk for monetary reward. Required participant number for each condition (*) was estimated with the effect size of impression manipulation in participant-level
regressions from a previous validation study with the same design ($R^2 > .65$, Todorov et al., 2013) so that the statistical power would reach .95.

**Materials.** I generated face stimuli with FaceGen (Singular Inversions, Toronto, Canada) from a data-driven model of impressions of competence (Todorov & Oosterhof, 2011). A data-driven model extracts the visual information used to form an impression without constraining the search to a priori set of facial features (Jack & Schyns, 2017; Todorov et al., 2011). In the FaceGen model, each face is a vector with 100 parameters in a 100-dimensional face space. A change in a parameter causes a holistic change in facial appearance, which is orthogonal to changes caused by other parameters. Fifty of the 100 parameters determine the shape of a face and the 50 other parameters determine its reflectance (i.e., texture and pigmentation). As I did in Part II, based on people’s ratings of many ($n \geq 300$) randomly generated faces on a single social impression (e.g., competence), one can calculate the contribution of each parameter to this impression and, consequently, one can model this impression as a linear vector in the statistical face space (Oosterhof & Todorov, 2008; Todorov & Oosterhof, 2011). With this method, Todorov et al. (2013) created and validated a model of competence impressions and a model of attractiveness impressions, among other models. In this paper, I used these two models. In Study 4a, I used the standard competence model (Todorov et al., 2013). In Studies 4b and 5–7, I used a new model created by subtracting the attractiveness from the competence model. Fig. 14 demonstrates how a synthetic face changes when it is manipulated by the models of competence, attractiveness, and their difference.
Fig. 14. A face manipulated by the competence model (top), the attractiveness model (middle), and the difference_{comp-attr} model (bottom). As the standard deviation (SD) units increase, the face is perceived as more competent (top), more attractive (middle), and more competent but not attractive (bottom).

Before applying the computational models to faces, I created 25 distinct, novel faces (see Appendix M for the original face images). These 25 faces were chosen from a random sample of 1,000 faces as those with the largest differences, based on the average Euclidean distance between the faces (Todorov et al., 2013). Because the process of
choosing distinct faces resulted in 25 atypical faces, the faces were scaled to more closely resemble the average face while still preserving the ratio of differences between them. Then, for each identity, I generated seven faces: each identity was projected at -3, -2, -1, 0, 1, 2, and 3SD on the dimension of the competence model. Each SD value represents the expected amount of change in competence ratings that would be caused by the corresponding change in the appearance of the face relative to the average face (Todorov et al., 2013). The complete set of stimuli consisted of 175 face images = 25 identities * 7 manipulation levels.

Procedure. Participants were randomly assigned to make judgments of either competence or attractiveness. The order in which the stimuli were presented was randomized. For each stimulus, the question asked was “How [trait] is this person?” presented with a 9-point scale, ranging from 1 (not at all [trait]) to 9 (extremely [trait]). The participants were blind to the social dimension on which the faces had been manipulated. Before the study began, each participant was told to rely on their “gut instinct” and not spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time for each trial.

To assess intrarater reliability, I added 25 repeated trials randomly chosen from the first 175 trials in each study. These extra 25 trials brought the total number of trials to 200. Using the ratings from the 25 repeated trial pairs, I calculated a correlational coefficient as a measure of test-retest reliability of each participant. The ratings from participants with zero or negative reliability were excluded, which resulted in responses of 15 participants with test-retest reliability >0 per impression. The interrater reliabilities of the impression ratings were high (competence: $\alpha = .91$, attractiveness: $\alpha = .97$).
Study 4a: Results and Discussion

To test whether competence and attractiveness impressions tracked the competence model manipulation, linear and quadratic regression models were fitted for the impression ratings. For the regressions, the impression ratings were averaged across participants (face-level analysis, \( n = 25 \) each) and across face identities (participant-level analysis, \( n = 15 \) each). The model fit was good across all models, showing that the judgments were well explained as a function of the manipulation level (Fig. 15, top).

The effect of the model manipulation on the competence and attractiveness ratings was consistent across face identities (Fig. 15, top; see Appendix O for the model fits for individual identities). The linear model explained more than 75% of the variance in the ratings (competence: \( R^2 = .83, F(1,173) = 824.30, P < .001 \); attractiveness: \( R^2 = .79, F(1,173) = 661.80, P < .001 \)). Although both competence and attractiveness ratings were explained by the competence manipulation, when we compared the coefficients from the regression models for competence and attractiveness ratings using multivariate models (Zellner, 1963), the face impression manipulation induced a bigger change in the competence (\( b_1 = 0.56 \)) than in the attractiveness ratings (\( b_1 = 0.50; Z = 3.20, P < .01 \)). This is not surprising given that the competence model had been built to capture the facial information underlying perceived competence, not attractiveness.
Fig. 15 The mean impression rating of competence (left) and attractiveness (right) as a function of the competence model (top) and the difference [competence – attractiveness] model (bottom) manipulation (Studies 4a – 4b).

For each impression, linear (solid line) and quadratic (dashed line) models were fitted for the ratings averaged across participants. $b_1$ is the unstandardized coefficient of the linear models. Error bars denote $\pm$SE.
The quadratic model explained more than 80% of the variance (competence: $R^2 = .86$, $F(2,172) = 517.0$, $P < .001$; attractiveness: $R^2 = .80$, $F(2,172) = 354.1$, $P < .001$). The quadratic models explained significantly more variance than the linear models did (competence: $F(1,172) = 37.21$, $P < .001$; attractiveness: $F(1,172) = 10.39$, $P < .01$). When we compared the coefficients from the regression models for competence and attractiveness ratings, again, the face impressions manipulation induced a bigger change in the competence than in the attractiveness ratings for both the quadratic (competence: $b_2 = -0.06$, attractiveness: $b_2 = -0.04$; $Z = 3.11$, $P < .01$) and linear terms (competence: $b_1 = 1.06$, attractiveness: $b_1 = 0.78$; $Z = 2.48$, $P < .05$).

The results were similar when the analysis was conducted at the level of participants (see Appendix N for detailed results). The linear models explained more than 50% of the variance in the ratings (competence: $R^2 = .54$, $F(1,103) = 122.3$, $P < .001$; attractiveness: $R^2 = .57$, $F(1,103) = 135.9$, $P < .001$). The quadratic models explained more than 55% of the variance (competence: $R^2 = .56$, $F(2,102) = 65.75$, $P < .001$; attractiveness: $R^2 = .58$, $F(2,102) = 69.66$, $P < .001$).

The results show that when facial cues of competence impressions are enhanced or attenuated, both competence and attractiveness impressions increase or decrease together. This covariance suggests that under natural circumstances, the visual components of competence and attractiveness impressions from faces are highly confounded with each other. This confounding relationship is consistent with the idea that the halo effect can explain competence impressions.

How is this finding related to gender biases in face-based impressions? It is important to note that the strong positive correlation between competence and
attractiveness impressions shown in previous research and Study 4a suggests that without facial attractiveness, competence impressions would significantly decrease, but especially for women. Feminine facial features lead to facial attractiveness, especially for female faces (e.g., Perrett et al., 1998, Said & Todorov, 2013). Given the overlap in the facial features between femininity and attractiveness impressions, removal of attractiveness would hurt the women more than men. That is, simply put, women might be perceived as competent to the extent that they are also attractive, whereas men might be perceived as competent regardless of whether they are attractive or not. I test these ideas in the rest of Part III. In the next section, I investigated what other ingredients may underlie the competence impressions besides attractiveness.

**Study 4b: Introduction**

The results of Study 4a show that facial attractiveness is a major ingredient of competence impressions. However, it is unclear whether there are other meaningful ingredients when attractiveness is not positively correlated with competence impressions. In Study 4b, I created a new model, i.e., the difference between the competence and attractiveness models (the difference \[\text{comp-attr} \] model), and applied this model to new faces. Theoretically, this model should force judgments of competence and attractiveness to be negatively correlated. However, the mapping between the model space and the psychological judgment space may not be linear (Oosterhof & Todorov, 2008; Walker & Vetter, 2016). To test how judgments of competence and attractiveness change as a function of the difference model, participants were asked to evaluate faces manipulated by this model on either competence or attractiveness.
Study 4b: Methods

Participants. A hundred-and-twenty-five online participants (72 males, 53 females) participated through Amazon MTurk for monetary reward. Required participant number for each condition ($n > 44$) was estimated with a medium effect size of impression manipulation in participant-level regressions ($R^2 = .20$) so that the statistical power would reach .90. I expected a medium effect size since removing attractiveness from the competence model (the resulting difference$_{[\text{comp-attr}]}$ model) should attenuate the effects of the model manipulation on judgments.

Materials. To create the difference$_{[\text{comp-attr}]}$ model, I subtracted each of the 100 parameters defining the attractiveness model from the 100 parameters defining the competence model. To the extent that the difference$_{[\text{comp-attr}]}$ model works, faces manipulated to be perceived as more competent should also be perceived as less attractive than faces manipulated to be perceived as less competent, or at the very least, as attractive as them.

The same 25 identities created for Study 4a were employed (Appendix M). Each identity was projected at -3, -2, -1, 0, 1, 2, and 3SD on the dimension of the difference$_{[\text{comp-attr}]}$ model, resulting in 7 faces per identity. As a result, as in Study 4a, the complete set of stimuli consisted of 175 face images = 25 identities * 7 manipulation levels.

Procedure. The procedure was identical to Study 4a except that faces were created using the difference$_{[\text{comp-attr}]}$ model. As in Study 4a, the ratings from participants with zero or negative reliability were excluded, which resulted in responses of 45 participants with test-retest reliability >0 per impression condition. The interrater reliabilities were high (competence: $\alpha = .90$, attractiveness: $\alpha = .95$).
Study 4b: Results and Discussion

To test whether competence and attractiveness impressions tracked the difference[comp-attr] model manipulation, linear and quadratic regression models were fitted for the impression ratings. For the regressions, the impression ratings were averaged across participants (face-level analysis, \( n = 25 \) each) and across face identities (participant-level analysis, \( n = 45 \) each). The model fit was good for the competence but not the attractiveness judgments, showing that only the competence judgments were well explained as a function of the manipulation level (Fig. 15, bottom). As the model manipulation level increased, ratings of competence increased too, but ratings of attractiveness did not.

The effect of the model manipulation on the competence and attractiveness ratings was consistent across face identities (Fig. 15, bottom; see Appendix O for the model fits for individual identities). However, the impression manipulation was far more impactful on the competence than on the attractiveness ratings. The linear models explained 49% of the variance in the competence ratings \( (F(1,173) = 163.00, P < .001) \) but only 6% of the variance in the attractiveness ratings \( (F(1,173) = 11.61, P < .001) \).

When we compared the coefficients from the regression models for competence and attractiveness ratings using multivariate models (Zellner, 1963), the face impressions manipulation induced a far bigger change in the competence \( (b_1 = 0.21) \) than in the attractiveness ratings \( (b_1 = -0.06; Z = 17.39, P < .001) \). This was true even when the sign of the coefficients in the attractiveness regression model was reversed to positive \( (Z = 5.16, P < .001) \) to match that of the competence regression
model. This finding shows that the difference model was indeed capable of varying the faces’ perceived competence without varying their attractiveness ratings too much. If anything, more competent-looking faces were perceived to be less attractive.

The quadratic model explained 66% of the variance in the competence ratings \( F(2, 172) = 165.5, P < .001 \) but only 13% of the variance in the attractiveness ratings \( F(2, 172) = 13.26, P < .001 \). The quadratic models explained more variance than the linear models did (competence: \( F(1, 347) = 101.32, P < .001 \); attractiveness: \( F(1, 347) = 16.81, P < .001 \)). When we compared the coefficients from the regression models for competence and attractiveness ratings, again, the face impressions manipulation induced a bigger change in the competence than in the attractiveness ratings for both the quadratic (competence: \( b_2 = -0.07 \), attractiveness: \( b_2 = -0.04; Z = 4.40, P < .001 \)) and linear terms (competence: \( b_1 = 0.79 \), attractiveness: \( b_1 = 0.23; Z = 8.17, P < .001 \)).

The results were consistent with the analysis conducted at the level of participants (see Appendix N for detailed results). The linear models explained 8% of the variance in the competence ratings \( F(1, 313) = 26.27, P < .001 \), but insignificant amount (< 1%) of the variance in the attractiveness ratings \( F(1, 313) = 1.36, P = .245 \). The quadratic models explained 11% of the variance in the competence ratings \( F(2, 312) = 18.32, P < .001 \), but insignificant amount (< 1%) of the variance in the attractiveness ratings \( F(2, 312) = 1.45, P = .237 \).

The results showed that when facial cues of competence impressions are enhanced by the difference\(_{[\text{comp-attr}]}\) model, competence impressions increased but attractiveness impressions decreased (participant-level analysis) or did not vary at all
(face-level analysis). This negative or null correlation between competence and attractiveness impressions contrasts with the high positive correlation between these impressions when facial cues of competence impressions were manipulated by the standard model (Study 4a). These results show that perceived competence can be meaningfully manipulated controlling for the halo effect of attractiveness.

**Study 5: Introduction**

Study 4b shows that competence impressions can be manipulated in the absence of the halo effect of attractiveness, suggesting that there are other visual ingredients of competence impressions besides attractiveness. Visual inspection of the difference model (Fig. 14, bottom) shows that as the faces increase in perceived competence, they express more confidence and look more masculine. This is consistent with prior research showing strong associations between competence impressions, confidence impressions, and gender (e.g., Spence et al., 1975; Vogel et al., 1972), as well as research showing high correlations between femininity and attractiveness (Said & Todorov, 2011). To formally test this hypothesis, participants were asked to evaluate faces – varying on perceived competence but not attractiveness – on either masculinity or confidence.

**Study 5: Methods**

*Participants.* Ninety-eight online participants (48 males, 49 females, and 1 other gender) participated through Amazon MTurk for monetary reward. Required participant number for each condition ($n > 44$) was estimated with a medium effect size of impression
manipulation in participant-level regressions ($R^2 = .20$) so that the statistical power would reach .90.

**Materials.** The same 25 identities created for Studies 4a and 4b were employed (Appendix M). Each identity was projected at -3, -2, -1, 0, 1, 2, and 3SD on the dimension of the difference_{comp-attr} model, resulting in 7 faces per identity. As a result, as in Study 4b, the complete set of stimuli consisted of 175 face images = 25 identities * 7 manipulation levels.

**Procedure.** The procedure was identical to the previous experiments except that participants were randomly assigned to make judgments of either confidence or masculinity. As in previous experiments, the ratings from participants with zero or negative reliability were excluded, which resulted in responses of 45 participants with test-retest reliability >0 per impression condition. The interrater reliabilities were high (confidence: $\alpha = .89$, masculinity: $\alpha = .91$).

**Study 5: Results and Discussion**

To test whether the confidence and masculinity impression ratings tracked the difference_{comp-attr} model manipulation, linear and quadratic regression models were fitted for the impression ratings. For the regressions, the impression ratings were averaged across participants (face-level analysis, $n = 25$ each) and across face identities (participant-level analysis, $n = 45$ each). The model fit was good across all models, showing the judgments were well explained as a function of the manipulation level (Fig. 16).
The effect of the model manipulation on the confidence and masculinity ratings was consistent across face identities (Fig. 16; see Appendix Q for the model fits for individual identities). The linear model explained more than 85% of the variance in the ratings (confidence: $R^2 = .88$, $F(1,173) = 1284.00, P < .001$; masculinity: $R^2 = .92$, $F(1,173) = 2045.00, P < .001$). The quadratic model explained more than 85% of the variance (confidence: $R^2 = .89$, $F(2,172) = 747.80, P < .001$; masculinity: $R^2 = .94$, $F(2,172) = 1337.00, P < .001$). The quadratic models explained more variance than the linear models did (confidence: $F(1,347) = 36.81, P < .001$; masculinity: $F(1,347) = 62.66, P < .001$).

The results were similar when the analysis was conducted at the level of participants (see Appendix P for detailed results). The linear model explained more than 45% of the variance in the ratings (confidence: $R^2 = .45$, $F(1,313) = 259.0, P < .001$; masculinity: $R^2 = .73$, $F(1,313) = 829.2, P < .001$). The quadratic model explained more than 45% of the variance (confidence: $R^2 = .46$, $F(2,312) = 133.3, P < .001$; masculinity: $R^2 = .74$, $F(2,312) = 443.5, P < .001$).

The results show that when facial cues of competence impressions are enhanced by the difference$_{\text{comp-attr}}$ model, both confidence and masculinity impressions increase. These relationships were expected from the visual inspection of the model (Fig. 14) and the previous literature, as discussed earlier. Competence and attractiveness impressions are not positively correlated in the faces used here, as shown in Study 4a. It follows that the variance in the competence impressions cannot be attributed to the halo effect of attractiveness, which is a significant natural confound of competence impressions. Thus, the results show that confidence and masculinity cues are important ingredients of
competence impressions: ingredients that cannot be explained as a byproduct of attractiveness.

**Fig. 16.** The mean impression rating of confidence (left) and masculinity (right) as a function of the difference_{comp-attr} model manipulation (Study 6). For each impression, linear (solid line) and quadratic (dashed line) models were fitted for the ratings averaged across participants. $b_1$ is the unstandardized coefficient of the linear models. Error bars denote ±SE.

**Study 6: Introduction**

Study 5 suggests that competence impressions, in the absence of the halo effect, are based on facial cues related to confidence and masculinity. The latter is strongly related to perceptions of gender and suggests the presence of gender biases in competence impressions. To directly test this hypothesis, I asked participants to categorize faces – manipulated on perceived competence – as male or female. I used both faces varied by the standard competence model and by the difference_{comp-attr} model. Given the high positive correlation between competence impressions and masculinity, I
expected that (1) participants would be more likely to categorize the faces as male, as the
level of model manipulation increases, irrespective of the model (i.e., the competence or
the difference\textsubscript{comp-attr} model) and (2) this effect would be accentuated for faces generated
by the difference\textsubscript{comp-attr} model. Specifically, controlling for attractiveness, faces
manipulated to be perceived as less competent would be more likely to be categorized as
female.

**Study 6: Methods**

*Participants.* Thirty-one online participants (22 males, 9 females) participated through
Amazon MTurk for monetary reward. Required participant number \((n > 26)\) was
estimated with a small to medium effect size of the main effects of the manipulation level
and the model type, as well as the interaction between the two \((f = .25)\) so that the
statistical power would reach .95.

*Materials.* I used both the face images manipulated by the competence model and the
face images manipulated by the difference\textsubscript{comp-attr} model. This created a combined pool
of 350 face image stimuli = 2 models \(*\) 25 identities \(*\) 7 manipulation levels.

*Procedure.* Participants were asked to make a forced choice of perceived gender for each
face. All participants were exposed to faces from both the competence and
difference\textsubscript{comp-attr} models. Two versions of the study with the same length were created:
half of the participants were presented with 88 competence model faces and 87
difference\textsubscript{comp-attr} model faces, while the other half were presented with 87 competence
faces and 88 difference\textsubscript{comp-attr} model faces. There was no overlap in the face images
between the two versions of the study.
The order in which the 175 chosen stimuli were presented was randomized. For each stimulus, the question asked was “What is the gender of this person?” presented with two options: “male” or “female.” Left and right arrow keys were used to indicate one or the other gender, and the two versions of the experiment had different gender-key response mappings. The participants were blind to the impressions on which the faces had been manipulated. Before the study began, each participant was told to rely on their “gut instinct” and not spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time.

To assess intrarater reliability of data, I added 25 repeated trials randomly chosen from the first 175 trials in each study. These extra 25 trials brought the total number of trials to 200. As in previous studies, the ratings from participants with zero or negative reliability were excluded, which resulted in responses of 30 participants with test-retest reliability >0.

Study 6: Results and Discussion

Overall, faces were more likely to be categorized as male than as female: the proportion of “male” responses to all faces ($n = 350$) averaged across participants was significantly higher than 0.5 ($M = 0.79$, $SD = 0.31$; $t(349) = 17.55$, $P < .001$). This may be mainly attributed to the fact that the faces were bald, which creates a strong bias to perceive faces as male faces. Nevertheless, as the competence manipulation level increased in both models, the categorization of faces as “males” increased too (Fig. 17).

To test whether the perceived gender of faces tracked the model manipulations, I conducted a $7 \times 2$ [manipulation level $\times$ model type] repeated measures ANOVA on the
proportion of “male” responses for each face. This analysis found that perceived gender varied as a function of both manipulation level and the type of model, as well as their interaction (Fig. 17). First, faces were more likely to be categorized as male when they were manipulated to look more competent, irrespective of the model type, as indicated by a main effect of the impression manipulation level \((F(6, 144) = 464.32, P < .001, \eta^2 = .88)\). Second, faces were more likely to be categorized as male when they were manipulated by the competence model than when they were manipulated by the difference[comp-attr] model, as indicated by a main effect of the impression model \((F(1, 24) = 796.55, P < .001, \eta^2 = .69)\). Third, the difference[comp-attr] model led to a much larger difference in the proportion of “male” categorization responses as a function of the manipulation level than the competence model did, as indicated by the interaction effect between the manipulation level and the impression model \((F(6, 144) = 291.93, P < .001, \eta^2 = .78)\).

This interaction effect reveals that the two models had differential effects on gender perception. When faces were varied by the standard competence model, most of the faces were likely to be categorized as male, despite the main effect of the manipulation level. This bias to perceive the faces as male could be attributed to the fact that they were all bald. However, once the attractiveness of the faces was subtracted from the faces manipulated by the competence model, as in the faces manipulated by the difference[comp-attr] model, gender categorization responses changed dramatically. Whereas faces manipulated to be perceived as competent (but not more attractive) were categorized as male, faces manipulated to be perceived as less competent (but not less attractive) were categorized as female. This effect shows that once the positive
covariance between attractiveness and competence impressions is visually removed, the variance in the masculinity cues becomes much more prominent in the faces.

![Graph showing the mean proportion of "male" responses as a function of the level of the difference](image)

**Fig. 17.** The mean proportion of “male” responses as a function of the level of the difference \(\text{comp(attr)}\) model (black line) and the competence model (grey line) manipulation. Sigmoid functions were fitted for the response averaged across participants. Error bars denote ±SE.

**Study 7: Introduction**

Studies 5 and 6 showed that when facial cues of competence impressions are enhanced, controlling for attractiveness, faces become more masculine. These results show that facial masculinity increases competence impressions, suggesting gender biases in these impressions. However, because I used synthetic faces, which tend to be categorized as male faces because of lack of strong gender cues, such as hair (Study 6), it is unclear whether the effects of masculinity cues work in the same way for male and
female faces. By adopting real-life face images, rather than computer-generated face images, Study 7 (a) extends the findings from Studies 4 – 6 to real-life images of faces, which are unambiguously categorized as male or female, and (b) tests for potential, differential effects of masculinity cues on competence impressions of male and female faces.

I manipulated photorealistic male and female faces using the difference[comp-attr] model. For male faces, I expected a monotonic increase in competence impressions, as faces were manipulated to be more masculine. For female faces, the predictions are less clear, as both empirical findings and computational work show that female faces are evaluated more negatively if their appearance is counterstereotypical (Sutherland et al., 2015). Part I evidenced this very effect where people evaluate impressions of women more negatively when their faces are less stereotypical. Given the strong gender stereotypes, it is possible that masculinity cues would have no effect or even a negative effect on competence impressions of female faces. Alternatively, while weak masculinity cues may increase these impressions, strong masculinity cues may decrease them.

**Study 7: Methods**

*Participants.* Two hundred and sixty-eight online participants (153 males, 115 females) participants through Amazon MTurk for monetary reward. Required participant number \(n > 122\) was estimated with the small effect size of the model manipulation on the competence rating from Study 4b \(R^2 = .08\) so that the statistical power would reach .90.

*Materials.* To apply a computational model to real-life face images, I transformed real face images using PsychoMorph (Tiddeman et al., 2001). First, I selected 10 male and 10
female real-life face images of self-identified Caucasian individuals (Appendix R) from a standardized face image set (DeBruine & Jones, 2017). Then, I created synthetic faces that represented extreme faces (-3 and 3SD, shown in the bottom row of Fig. 14) generated by the difference model. Next, I manipulated the 20 real-life face images along the continuum of the difference between the two extreme face images in shape, color, and texture. The transformation procedure allowed me to generate photo-realistic images that varied in facial information captured by the difference model (Fig. 18). On the most extreme ends, each face image was transformed 25% away from the original face. The manipulation magnitude was identical for the intervals between the 4 facial variations: the final face images were transformed -25.00%, -8.33%, 8.33%, and 25.00% away from the initial images. The complete set of stimuli consisted of 80 face images = 2 (male or female) * 10 identities (male or female) * 4 manipulation levels.
Fig. 18. A sample of real-life face images manipulated by the difference \( \text{comp-attr} \) model. As the manipulation increases, the faces are perceived as more competent. The unit of manipulation represents the extent to which the shape and reflectance of each original face image was transformed towards or away from the extreme ends of the model (i.e., -3 and 3SD faces in Fig. 14, bottom).

Procedure. Participants were randomly assigned to make judgments of competence of either male (male rating condition) or female faces (female rating condition). The order in which the stimuli were presented was randomized with the constraint that face images generated from the same original face identity were never consecutively shown. For each stimulus, the question asked was “How competent is this person?” presented with a 9-point scale, ranging from 1 (not at all competent) to 9 (extremely competent). Before the study began, each participant was told to rely on their “gut instinct” and not spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time to respond.
To assess intrarater reliability, I added 10 repeated trials randomly chosen from the first 40 trials in each study. These extra 10 trials brought the total number of trials to 50. Using the ratings from the 10 repeated trial pairs, I calculated a correlational coefficient as a measure of test-retest reliability of each participant. As in previous studies, the ratings from participants with zero or negative reliability were excluded, which resulted in responses of 125 participants with test-retest reliability > 0 per face gender. The interrater reliabilities of the impression ratings were high (male faces: $\alpha = .94$, female faces: $\alpha = .95$).

**Study 7: Results**

To test whether competence impressions tracked the difference model manipulation, linear and quadratic regression models were fitted for the competence ratings. For the regressions, the ratings were averaged across face identities (participant-level analysis, $n = 125$ per gender; Fig. 19).

The linear model explained a significant amount of variance in the ratings of male faces, but not in the ratings of female faces (male faces: $R^2 = .02, F(2,498) = 11.94, P < .001$; female faces: $R^2 = .01, F(2,498) = 2.40, P = .121$). The quadratic model explained a significant amount of variance in the ratings of faces of both genders (male faces: $R^2 = .02, F(2,497) = 6.04, P < .01$; female faces: $R^2 = .02, F(2,497) = 4.29, P < .05$). However, for male faces, the quadratic model fit was not significantly better than the linear model fit ($F(2,497) = 0.16, P = .69$).

Although the observed effect sizes of the model manipulation were significant, they were relatively small. This may be due to several factors: the difference...
model does not directly manipulate attractiveness facial cues, which play a big role in competence impressions; the difference_{comp-attr} model was derived from the ratings of synthetic, not real-life, face images and therefore might have had a smaller effect when applied to real-life faces; real-life face images are more distinctive than synthetic face images with individual face identities, adding to unexplained variance; participants regarded face images originated from the same identity as identical and therefore gave the same rating across these images before closely examining the facial cues.

Nevertheless, the results show that when facial cues of competence impressions are enhanced using the difference_{comp-attr} model, competence impressions increased, although the effects were different for male and female faces. Whereas male faces became more competent looking through the increasing manipulation levels (Fig. 19, left), female faces became more competent looking only up to a point after which their perceived competence decreased (as shown in the significant quadratic fit of the ratings; Fig. 19, right). These results show that (a) perceived competence of real-life male and female faces can be meaningfully altered controlling for the halo effect of attractiveness, (b) male faces receive incremental benefit as masculinity/confidence cues increase, and (c) female faces receive limited benefit as masculinity/confidence cues increase and, in fact, are perceived as less competent once these cues become too strong (I obtained similar results using a competence model orthogonal to the attractiveness model; see Appendix S for details). The latter finding is consistent with prior research showing that counterstereotypical (e.g., dominant) female faces are evaluated negatively (e.g., Sutherland et al., 2015; Studies 1 - 2 of the current thesis) and theoretical frameworks
posing that women are evaluated positively only when they fit narrowly defined female gender norms (e.g., Glick & Fiske, 1996).

**Fig. 19.** The mean competence ratings of real-life male (left) and female faces (right) as a function of the difference$^{\text{[comp-attr]}}$ model manipulation (Study 7). For each impression, linear (gray line) and quadratic (black line) models were fitted for the ratings averaged across faces. $b_1$ and $b_2$ are unstandardized coefficients of the models. Error bars denote ±SE.

**Part III: Interim Conclusion**

In Part III, I found that multiple underlying features of competence impressions from faces using data-driven computational models of impressions. First, I showed that competence impressions naturally correlate with facial attractiveness (Study 4a). When the variance in competence impressions could not be attributed to facial attractiveness, more competent faces were perceived as more confident and masculine (Study 4b and 5) and were more likely to be categorized as male (Study 6). When the same
confidence/masculinity cues were increased on real-life male and female faces, it monotonically increased the competence impressions of male faces, but not those of female faces (Study 7). Rather, the confidence/masculinity cues increased female faces only to a point and when the cues were too strong, female faces were perceived as less competent.

The last study (Study 7) is noteworthy since it directly links Part I and II with Part III. Because competence is a positive trait (Study 4; Fiske et al., 2002) and masculine facial features are at the basis of competence impressions (Studies 5–6), it is difficult for a woman to acquire competence impressions. It would require her to have conflicting sets of facial cues: facial cues related to overall positive impression (stereotypical, therefore feminine, facial looks) and facial cues related to masculinity impression (counterstereotypical, therefore masculine, facial looks). This highlights the unfair nature of gender biases in competence impressions.

In sum, consistent with the idea that competence impressions have multiple components, I found that impressions of attractiveness, confidence, and masculinity are at the basis of face-based competence impressions. Crucially, the findings suggest the existence of strong gender biases that shape competence impressions. Specifically, men are more likely to be perceived as competent, an effect that was not readily apparent from a model of competence impressions that did not control for facial attractiveness. These findings have significant implications for social perception research and social fairness.

First, the results illustrate a useful way to overcome natural confounds in visual social perception (Todorov et al., 2013). Prior research (Eagly et al., 1991) and Study 4a showed a strong positive relationship between judgments of attractiveness and
competence. Generally, attractiveness correlates with many other social judgments (Dion et al., 1972; J. R. Graham et al., 2017; Todorov et al., 2013; Webster & Driskell, 2015), making it difficult to test whether the effects of various social impressions on behaviors are due to specific impressions, such as trustworthiness and competence, or the halo effect of attractiveness. However, within the present computational framework, it is straightforward to control for facial attractiveness. I illustrated this method by simply subtracting the model of attractiveness from the competence model. This is the strongest manipulation of controlling for attractiveness, because it forces the two judgments, as shown in Study 4b, to be negatively or uncorrelated. It is also possible to orthogonalize two different models (see Appendix T for sample face images), but the orthogonality between the two models in the face model space does not guarantee that the judgments of faces generated by the two models would be orthogonal. The orthogonality between the two judgments requires an empirical test. In addition to controlling for undesirable confounds, these methods have the potential to reveal visual ingredients of impressions that are not readily discernible from the models of these impressions. As shown in Studies 5 – 7, once controlling for attractiveness, it was readily apparent and later validated that competence impressions rely on masculinity cues.

Second, the current findings are consistent with a body of literature in psychology and gender studies that has shown a strong link between competence impressions and male-ness. People report that they as well as other people (“society”) believe that men are more competent than women are (Bem, 1974; Prentice & Carranza, 2002; Spence et al., 1975; Vogel et al., 1972). This gender bias in perceived competence seems to emerge early in development. Girls as young as 6 years old, unlike boys of the same age, think
that their gender on average lacks a high level of intellectual competence (Bian, Leslie, & Cimpian, 2017). These biased perceptions most likely exacerbate gender inequality in various settings: women are more likely to be discriminated in professional settings where they are perceived as competent (Hagen & Kahn, 1975; Rudman & Phelan, 2008), and they are more likely to face obstacles when entering or staying in a field that requires intellectual competence (Leslie, Cimpian, Meyer, & Freeland, 2015; Moss-Racusin et al., 2012; Shen, 2013).

It is noteworthy that certain feminine facial features, especially feminine face shape, make both male and female faces more attractive (Little et al., 2001; Penton-Voak et al., 1999; Perrett et al., 1998; Rhodes et al., 2000; Said & Todorov, 2011, but see Rhodes, 2006). Thus, one may think that women, especially those with attractive or feminine looks, may benefit from the halo effect of attractiveness when a stranger judges their competence. However, the current findings suggest that masculinity cues play a crucial role in competence impressions, hurting women to the extent that they possess feminine (i.e., less masculine) looks. Moreover, when women possess strong masculine looks, women may be discriminated because of the counterstereotypical nature of these looks, as shown in Sutherland et al. (2015) and the findings in Parts I (Study 1), II (Study 2), and III (Study 7). These gender biases in competence impressions are masked because the attractiveness resulted from feminine facial features contributes to competence impressions. By effectively removing the influence of attractiveness on competence impressions, Studies in Part III uncover strong gender biases in these impressions. These biases are particularly alarming as intuitive competence judgments have powerful effects on leadership selection (Antonakis & Eubanks, 2017).
Conclusion

The gender of an individual shapes our impressions of them (Mattarozzi et al., 2015; McAleer, Todorov, & Belin, 2014; Said & Todorov, 2011; Sutherland et al., 2015; 2016). This results in differential treatment of men and women in diverse settings, including academia (Leslie et al., 2015; Moss-Racusin et al., 2012; Shen, 2013; Steinpreis, Anders, & Ritzke, 1999; Wennerås & Wold, 1997) and industry (Hagen & Kahn, 1975; Moss-Racusin et al., 2010; Phelan & Rudman, 2010; Rudman, 1998; Rudman & Phelan, 2008). Given its significant and expansive effect, the mechanism by which gender affects impressions deserves a thorough investigation (Oosterhof & Todorov, 2008; Sutherland et al., 2013; 2015; Wolffhechel et al., 2014).

Deleterious gender biases in facial impressions. The current work investigated how face gender affects person impressions. The findings suggest deleterious gender biases in face-based person impressions. These biases serve as direct evidence against the claim that women are no longer oppressed. The findings here show that women are at a disadvantage in the beginning of almost any interaction, namely, the moment when others form impressions of them. First, impressions of women are simplified; women are negatively evaluated to the extent that their looks are atypical (Part I), although the same facial information is used to form impressions of both men and women (Part II). Second, women may be perceived as less competent when their looks are feminine, especially when their faces are not attractive (Part III). Given how first impressions are prioritized during person evaluation over information acquired later (Asch, 1946), the current findings may partly explain the prevalent inequality between men and women.
Narrow boundary conditions of positive impressions of women. When considering the findings of Parts I – III together, we notice how narrowly the boundary conditions for positive impressions of women are defined. That is, everything else being equal, it is more difficult for a woman to be perceived positively than a man. In the case of competence, which is an unambiguously positive trait, a face needs to possess both gender-typical looks to induce overall positivity of impressions (shown in Parts I and II) as well as masculine looks to induce competence impressions (shown in Part III). The findings reported here show that not only the latter (masculine looks) is less likely to happen for women in a pure probabilistic sense, but also that these two looks (typical vs. masculine looks) intrinsically conflict for women. The current findings support and extend theoretical work that suggests that women have narrower boundary conditions for positive social perception compared to men (e.g., Glick & Fiske, 1996; Rudman, 1998).

Importance of the target category and the perceiver characteristics in social perception. Aside from the implications regarding gender in/equality, the current work sheds light on the process of person perception in several important ways. First, the current work highlights the significance of social categories in visual social perception. Across studies, male and female faces were evaluated in meaningfully different ways (e.g., more simplified impressions for women, limited or no benefit of masculinity cues on female competence impressions), corroborating previous research (e.g., Sutherland et al., 2015). Further, facial features related to gender were used to form a meaningful impression (e.g.,
masculinity, a gender-related cues, leads to competence impressions, thereby favoring men), corroborating previous research (e.g., Hess et al., 2000).

Second, the current work highlights the significant role of the perceiver in visual social perception. For instance, when perceivers were more willing to endorse gender stereotypes, they showed more simplified impressions of women (Study 1b). The effect of perceiver characteristics on social perception has been shown to be crucial; the mental representation of a social group, for example, can be either positive or negative depending on the level of participants’ prejudice against the group (e.g., Brown-Iannuzzi, Dotsch, Cooley, & Payne, 2017; Lei & Bodenhausen, 2017). Additionally, the current findings support theoretical work that posits that impressions of target individuals can be intercorrelated to various degrees depending on the perceiver’s preconception of the group (e.g., a perceiver who believes more attractive females are less dominant and vice versa will show a strong negative intercorrelation between attractiveness and dominance for female faces; Secord, 1958; Secord & Berscheid, 1963; Stolier, Hehman, & Freeman, 2017).

All in all, using behavioral experiments, data-driven modeling, and face morphing, in this thesis, I addressed the question of how face gender influences person impressions and how gender biases play out in the process of impression formation. The current thesis shows that the target category and perceiver characteristics are crucial in unraveling the rich and complicated processes of social perception.
References


Appendix A

Intercorrelations between impressions from male and female faces in Studies 1a – 1c.

I conducted Pearson correlational analyses between all pairs of impression ratings within each face gender. In each cell, the color (dark: strong correlation, light: weak correlation) and the angle of an ellipse (acute rotation angle: positive correlation, obtuse rotation angle: negative correlation) represents the direction and strength of the correlation between a pair. The rows and columns of each matrix are clustered according to the levels of correlations between impressions ratings.

(continued)
Appendix B

Principal component analysis results in Studies 1a – 1c.

The magnitude of the bars indicates the loading strength of traits on the first (left) and the second (right) principal components for male and female faces in Studies 1a – 1c. The traits are sorted by their loadings for male faces on each component.
Appendix C

Trait terms used to measure individual raters’ levels of gender stereotype endorsement in Study 1b.

Each of the questions read “How do the average man and the average woman compare with each other on how [trait term] they are?” Emboldened text denotes six traits that elicited a high level of item-whole correlation ($r > .6$). These traits were used to construct the GSE index. See the main text for details.

<table>
<thead>
<tr>
<th>Trait Gender</th>
<th>Valence</th>
<th>Trait Item</th>
<th>Item-Whole Correlation (Correlation of each item with the rest of the questionnaire)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stereotypical</td>
<td>Positive</td>
<td>dominant</td>
<td>.66</td>
</tr>
<tr>
<td>Stereotypical</td>
<td>Positive</td>
<td>competitive</td>
<td>.61</td>
</tr>
<tr>
<td>Stereotypical</td>
<td>Positive</td>
<td>quantitative</td>
<td>.47</td>
</tr>
<tr>
<td>Stereotypical</td>
<td>Positive</td>
<td>analytical</td>
<td>.41</td>
</tr>
<tr>
<td>Male Traits</td>
<td>Negative</td>
<td>aggressive</td>
<td>.64</td>
</tr>
<tr>
<td>Male Traits</td>
<td>Negative</td>
<td>egotistical</td>
<td>.59</td>
</tr>
<tr>
<td>Male Traits</td>
<td>Negative</td>
<td>boastful</td>
<td>.57</td>
</tr>
<tr>
<td>Male Traits</td>
<td>Negative</td>
<td>hostile</td>
<td>.57</td>
</tr>
<tr>
<td>Male Traits</td>
<td>Negative</td>
<td>arrogant</td>
<td>.53</td>
</tr>
<tr>
<td>Male Traits</td>
<td>Negative</td>
<td>cynical</td>
<td>.26</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Positive</td>
<td>nurturing</td>
<td>.64</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Positive</td>
<td>sensitive</td>
<td>.64</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Positive</td>
<td>artistic</td>
<td>.23</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Positive</td>
<td>intuitive</td>
<td>.11</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Stereotypical</td>
<td>emotional</td>
<td>.62</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Negative</td>
<td>subordinate</td>
<td>.54</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Negative</td>
<td>nagging</td>
<td>.53</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Negative</td>
<td>servile</td>
<td>.44</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Negative</td>
<td>whiny</td>
<td>.41</td>
</tr>
<tr>
<td>Female Traits</td>
<td>Negative</td>
<td>gullible</td>
<td>.34</td>
</tr>
</tbody>
</table>
Appendix D

The distribution of raters’ responses to gender stereotype endorsement questions in Study 1b.

In every question, raters showed a significant bias away from the middle score (blue dashed line) in the expected direction consistent with gender stereotypes ($t_s > 3.87, P < .001$).
Appendix E

Analysis of the effects of raters’ gender stereotype endorsement.

When testing for the effect of the raters’ GSE on the impression, I conducted three additional analyses on the structure of impressions using the total score of GSE (the total GSE score; the score that included all items in the GSE questionnaire) in addition to the main analyses reported in the main text (Study 1b). In the main analyses, I used the GSE index, which included only the six items with the highest inter-item reliability (see the main text for details). The total GSE score had a possible range of [20, 180], in which a higher score indicates a higher level of stereotype endorsement. Again, as the GSE index, the total GSE score was higher than the absolute middle score (i.e., 100; $M = 121.91$, $SD = 17.09$; $t(451) = 27.25$, $P < .001$), indicating that participants on average endorsed gender stereotypical beliefs. Again, male raters showed a higher level of gender stereotype endorsement ($M = 126.04$, $SD = 19.21$) than female raters ($M = 117.88$, $SD = 13.61$; $t(399.1) = 5.20$, $P < .001$).

The three additional analyses yielded consistent results as the analysis reported in the main text. In the first analysis, I selected raters with high and low total GSE scores by selecting the top and bottom 33% of the raters. A participant-level analysis (e.g., a hierarchical regression) was not executed because it required ratings on all the dimensions by each rater, which was avoided to prevent participant fatigue. To examine the role of the rater GSE, I conducted two analyses (i.e., the comparison of the level of correlations across impression ratings and the PCA on the impression ratings). First, I calculated correlational coefficients across impression ratings separately for the high- and the low-GSE participants. I then conducted a 2 [face gender] × 2 [GSE levels] repeated measures ANOVA on the absolute values of the inter-trait correlational coefficients.

In the ratings of both high and low GSE raters, PC1 explained more variance in female face impressions than in male face impressions (51.82% vs. 50.23% in the high GSE group, 39.28% vs. 35.86% in the low GSE group). The 2 [face gender] × 2 [GSE levels] repeated measures ANOVA on the absolute values of the intercorrelational coefficients across traits yielded a significant main effect of GSE ($F(1,90) = 29.22$, $P < .001$, $\eta^2_G = .09$), where the responses from the high GSE individuals showed a higher level of correlations between social traits ratings of faces ($M = 0.44$, $SD = 0.25$) than the low GSE individuals ($M = 0.30$, $SD = 0.19$). Neither the main effect of face gender nor the interaction effect was significant. These
findings indicate that those who strongly endorsed gender stereotypes showed less differentiated impressions than those who did not strongly endorse these stereotypes, regardless of face gender.

In the second analysis, I divided raters based on median split, which yielded consistent results. In the ratings of both raters with higher or lower total GSE scores, PC1 explained more variance in female face impressions than in male face impressions (62.50% vs. 58.33% in the higher-GSE group, 56.28% vs. 54.32% in the lower-GSE group), although this difference was smaller in the lower-GSE group. The 2 [face gender] \times 2 [GSE levels] repeated measures ANOVA for the absolute values of the intercorrelational coefficients across traits yielded a significant main effect of GSE (F(1,90)=31.16, P<.001, \eta^2_G=.02), where the responses from the higher-GSE individuals showed a higher level of correlations between trait ratings of faces (M=0.56, SD = 0.24) than the lower-GSE individuals (M=0.49, SD = 0.22). There was a significant main effect of face gender (F(1,90) = 4.31, P < .05, \eta^2_G = .01), where female impressions showed a higher level of correlations between trait ratings of faces (M = 0.55, SD = 0.22) than male impressions (M = 0.51, SD = 0.24). The interaction effect between face gender and GSE level was not significant.

In the third analysis, I repeated the two main analyses in the main text (i.e., the correlational analyses and PCAs), using responses of a subset of raters, which I varied, this time, according to their total GSE score, not the GSE index (see the main text for the results with the GSE index). Overall, as the rater GSE score increased, the correlations between impression ratings increased too both for male and female faces (Appendix F for figures), but the relationship was nonlinear. Although the linear regression model suggested a meaningful relationship between the total GSE score and the level of impression intercorrelation for the ratings of both genders (male faces: \( R^2 = .11, F(1,42) = 5.03, P < .05 \); female faces: \( R^2 = .05, F(1,42) = 2.40, P = .130 \)), it accounted for far less variance than the quadratic model (male faces: \( R^2 = .91, F(2,41) = 198, P < .001 \); female faces: \( R^2 = .83, F(2,41) = 99.54, P < .001 \)). The difference between the absolute correlation coefficients for male and female face ratings was significant in the subset of raters with a mean total GSE score ranging from 110 to 119, as well as in the subset of raters with a mean total GSE score ranging from 135 to 137, as shown in the grey areas in Appendix F (left figure). That is, as the results reported in the main text, raters who more strongly endorsed gender stereotypes were more likely to show less differentiated impressions of female faces than of male faces, but this effect was not present for the raters with the strongest gender stereotypes. Correspondingly, the amount of variance
explained by PC1 in the ratings followed the same quadratic pattern of change across the total GSE score (Appendix F, right figure).
Appendix F

The level of intercorrelations among impressions (A) and the amount of variance in the impressions explained by PC1 (B) as a function of the raters’ total gender stereotype endorsement score in Study 1b.

Each data point was calculated from a subset of raters with sliding windows on the level of rater GSE ($n_p \geq 126$ for each subset). The $x$ value is the middle point of the sliding window. The shades surrounding data points denote $\pm$SE, and the rectangular shades denote a significant difference between the intercorrelations for male and female faces (light grey: $P < .05$, dark grey: $P < .01$). The semitransparent curves show quadratic fits of the data from the ratings of male (lighter grey) and female faces (darker grey).
Appendix G

A sample of randomly generated female (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.
Appendix H

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 (lighter grey triangles) and female faces (darker grey circles) for the trustworthiness and dominance ratings, respectively. $b_1$ is the unstandardized coefficient of the linear models.
Appendix I

Similarities between gender-specific trustworthiness (“T”) and dominance models (“D”).

Similarities between models are represented with angles (e.g., 0 rad when \( r = 1 \), \( \pi/2 \) rad when \( r = 0 \), \( \pi \) when \( r = -1 \)). Each model was built on the ratings of either only male faces (“M”, solid lighter grey line) or only female faces (“F”, solid darker grey line).

Below are the correlations between gender-specific trustworthiness and dominance models. Numbers indicate pairwise Pearson coefficients across gender-specific impression models.

<table>
<thead>
<tr>
<th>Social Trait Model</th>
<th>Male models</th>
<th>Female models</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>D</td>
</tr>
<tr>
<td>Male</td>
<td>1.00</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>- .16</td>
<td>1.00</td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>- .58**</td>
<td>- .44**</td>
</tr>
<tr>
<td>Dominance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trustworthiness</td>
<td>- .14</td>
<td>.85**</td>
</tr>
<tr>
<td>Dominance</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

*\( P < .001 \).
Appendix J

Interrater reliabilities of ratings of computer-generated faces manipulated by trustworthiness and dominance models in Study 3a.

<table>
<thead>
<tr>
<th>Social trait model and original face gender</th>
<th>Cronbach’s $\alpha$ based on face ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male faces manipulated with male trustworthiness model</td>
<td>.97</td>
</tr>
<tr>
<td>Male faces manipulated with male dominance model</td>
<td>.96</td>
</tr>
<tr>
<td>Face $\times$ Model</td>
<td>male dominance model</td>
</tr>
<tr>
<td>Gender-Congruent Faces</td>
<td>Female faces manipulated with female trustworthiness model</td>
</tr>
<tr>
<td></td>
<td>Female faces manipulated with female dominance model</td>
</tr>
<tr>
<td></td>
<td>Male faces manipulated with female trustworthiness model</td>
</tr>
<tr>
<td></td>
<td>Male faces manipulated with female dominance model</td>
</tr>
<tr>
<td>Face $\times$ Model</td>
<td>female dominance model</td>
</tr>
<tr>
<td>Gender-Incongruent Faces</td>
<td>Female faces manipulated with male trustworthiness model</td>
</tr>
<tr>
<td></td>
<td>Female faces manipulated with male dominance model</td>
</tr>
</tbody>
</table>
Appendix K

Twenty-five female (A) and twenty-five male (B) face identities used in the validation of the data-driven, computational models in Study 3a and Study 3b.

Below are faces that were randomly generated by a statistical face model with the constraint to be maximally distinctive from each other, selected from a set of 1000 randomly generated faces (25 per gender). Face stimuli for Study 3a were generated based on these faces.
Below are faces that were randomly selected from a standardized face stimulus set (25 per gender; DeBruine & Jones, 2017). Face stimuli for Study 3b were generated based on these faces.
Appendix L

Interrater reliabilities of ratings of real-life faces manipulated by trustworthiness and dominance models in Study 3b.

<table>
<thead>
<tr>
<th>Face × Model</th>
<th>Social trait model and original face gender</th>
<th>Cronbach’s $\alpha$ based on face ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender-Congruent Faces</td>
<td>Male faces manipulated with male trustworthiness model</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>Male faces manipulated with male dominance model</td>
<td>.93</td>
</tr>
<tr>
<td>Face × Model</td>
<td>Female faces manipulated with female trustworthiness model</td>
<td>.81</td>
</tr>
<tr>
<td></td>
<td>Female faces manipulated with female dominance model</td>
<td>.92</td>
</tr>
<tr>
<td></td>
<td>Male faces manipulated with female trustworthiness model</td>
<td>.83</td>
</tr>
<tr>
<td></td>
<td>Male faces manipulated with female dominance model</td>
<td>.91</td>
</tr>
<tr>
<td>Gender-Incongruent Faces</td>
<td>Male faces manipulated with male trustworthiness model</td>
<td>.85</td>
</tr>
<tr>
<td>Face × Model</td>
<td>Female faces manipulated with male dominance model</td>
<td>.92</td>
</tr>
</tbody>
</table>
Appendix M

The 25 original face identities used to create new faces varied by the competence model (Study 4a) and the difference [comp-attr] model (Studies 4b and 5 – 7).

Following the same procedure described in Appendix H (i.e., synthetic face stimuli for Study 3a), these faces were randomly generated by a statistical face model with the constraint to be maximally distinctive from each other, selected from a set of 1000 randomly generated faces.
Appendix N

Participant-level mean impression ratings of competence (left) and attractiveness (right) as a function of the competence model (top) and the difference $c_{\text{comp-attr}}$ model (bottom) manipulation (Studies 4a and 4b).

Linear models (the first four plots) were fitted for the ratings of each participant averaged across face identities (gray lines) and for the ratings averaged across participants and face identities (black line). Quadratic models (the next four plots) were fitted for the ratings of each participant averaged across face identities (gray lines) and for the ratings averaged across participants and face identities (black line). Error bars denote ±SE.

(continued)
Standard Competence Model (Study 4a), Linear Fit

Difference \(_{(\text{Competence} - \text{Attractiveness})}\) Model (Study 4b), Linear Fit

Competent

Attractive
Standard Competence Model (Study 4a), Quadratic Fit

Difference_{Competence – Attractiveness} Model (Study 4b), Quadratic Fit
Appendix O

Amount of variance of the competence (left) and attractiveness (right) ratings explained by the linear (grey bars) and quadratic models (black bars) for each face identity.

The face identities were manipulated by the competence model (top; Study 4a) or the difference[comp-attr] model (bottom; Study 4b).
Appendix P

Participant-level mean impression ratings of confidence (left) and masculinity (right) as a function of difference_{comp-attr} model manipulation (Study 5).

Linear models (top) were fitted for the ratings of each participant averaged across face identities (gray lines) and for the ratings averaged across participants and face identities (black line). Quadratic models (bottom) were fitted for the ratings of each participant averaged across face identities (gray lines) and for the ratings averaged across participants and face identities (black line). Error bars denote ±SE.

(continued)
Difference in Competence – Attractiveness Model (Study 5), Linear Fit

Confident

Masculine

Difference in Competence – Attractiveness Model (Study 5), Quadratic Fit

Confident

Masculine
Appendix Q

Amount of variance of the confidence (left) and masculinity (right) ratings explained by the linear (grey bars) and quadratic models (black bars) for each face identity.

The face identities were manipulated by the difference\textsubscript{[comp-attr]} model (Study 5).
Appendix R

Face identities used to create new faces varied by the difference_{comp-attr} model (Study 7) and the orthogonal_{comp⊥attr} model (Appendix P).

Below are 10 male (two top rows) and 10 female original face identities (two bottom rows) used to create new face images varied by the difference_{comp-attr} (Study 7) and orthogonal_{comp⊥attr} model (Appendix P).

The faces were chosen from London Face Set (DeBruine & Jones, 2017).
Appendix S

Supplemental results with faces manipulated by the standard competence model orthogonal to attractiveness.

I conducted an additional study, in which I used a competence model orthogonal to the attractiveness model to manipulate real-life images of male and female faces. This study is a replication of Study 7, using an orthogonal rather than a difference model, which is negatively correlated with the attractiveness model. As explained in the main text, the difference model provides a stronger control for the attractiveness confound. Nevertheless, it is important to show that the results are converging irrespective of the type of control for the attractiveness confound.

Methods

Participants. Two hundred and seventy-six online participants (160 males, 116 females) were recruited through Amazon MTurk, and participated for monetary reward. As in Study 7, required participant number \((n > 122)\) was estimated with the small effect size of the model manipulation on the competence rating from Study 4b \((R^2 = .08)\) so that the statistical power would reach .90.

Materials. I created and used the model of competence impressions that is orthogonal to the model of attractiveness impressions (the orthogonal\([\text{comp}_\perp \text{attr}]\) model, see Appendix T for the visualization of the model). When a face is varied by this model, across the face variants, the variance in the facial information related to attractiveness is statistically controlled. Specifically, the model consisted of parameters that are the residuals of the linear regression model in which the parameters of the competence impressions model were predicted by the parameters of the attractiveness model. As in Study 7, to apply a computational model to real-life face images, I transformed real face images (Appendix O) using PsychoMorph (Tiddeman et al., 2001; see the main text for details). As in Study 7, the final face images were transformed -25.00%, -8.33%, 8.33%, and 25.00% away from the initial images. The complete set of stimuli consisted of 80 face images = 2 (male or female) * 10 identities (male or female) * 4 manipulation levels.

Procedure. Participants were randomly assigned to make judgments of competence of either male (male rating condition) or female faces (female rating condition). The order in which the stimuli were presented was randomized with the constraint that face images generated from the same original face identity were never consecutively shown. For each stimulus, the question asked was “How competent is this person?”
presented with a 9-point scale, ranging from 1 (not at all competent) to 9 (extremely competent). Before the study began, each participant was told to rely on their “gut instinct” and not spend too much time on each face, and that there were no right or wrong answers. Participants were given unlimited time to respond.

To assess intrarater reliability, I added 10 repeated trials randomly chosen from the first 40 trials in each study. These extra 10 trials brought the total number of trials to 50. Using the ratings from the 10 repeated trial pairs, I calculated a correlational coefficient as a measure of test-retest reliability of each participant. The ratings from participants with zero or negative reliability were excluded, which resulted in responses of 125 participants with test-retest reliability >0 per face gender. The interrater reliabilities of the impression ratings were high (male faces: $\alpha = .94$, female faces: $\alpha = .94$).

Results

To test whether competence impressions tracked the orthogonal [$comp \perp atr$] model manipulation, linear and quadratic regression models were fitted for the competence ratings. For the regressions, again, as in Study 7, the ratings were averaged across face identities (participant-level analysis, $n = 125$ per gender).

The linear model explained a significant amount of variance in the ratings of male faces, but not in the ratings of female faces (male: $R^2 = .02$, $F(1,498) = 9.77$, $P < .01$; female: $R^2 = .01$, $F(1,498) = 3.60$, $P = .058$). The quadratic model explained a significant amount of variance in the ratings of male faces, but not in the ratings of female faces (male: $R^2 = .02$, $F(1,497) = 4.92$, $P < .01$; female: $R^2 = .01$, $F(1,497) = 2.72$, $P = .067$). For male faces, the quadratic model fit was not significantly better than the liner model fit ($F(1,497) = 0.096$, $P = .756$). Although analyses of the ratings of female faces did not reach the level of statistical significance, the findings are consistent with the results from Study 7: when facial cues of competence impressions are enhanced using a face model controlling for attractiveness, male faces receive incremental benefit, but female faces receive limited benefit as masculinity/confidence cues increase and, in fact, are perceived as less competent once these cues become too strong.
Appendix T

A face manipulated by the competence model that is orthogonal to the attractiveness model.

Across the standard deviation (SD) units of the model, the variance in the facial information related to attractiveness is statistically controlled. The model consists of parameters that are the residuals of the linear regression model in which the parameters of the competence impressions model are predicted by the parameters of the attractiveness model.
Appendix U

Real-life faces manipulated by the orthogonal $\{\text{comp}, \text{attr}\}$ model.

Real-life face images were manipulated using the competence model that is orthogonal to the attractiveness model (Appendix T). The unit of manipulation represents the extent to which the shape and reflectance of each original face image was transformed towards or away from the extreme ends of the model (i.e., -3 and 3SD faces in Appendix T).

Competence $\perp$ Attractiveness
Appendix V

The mean impression ratings of competence of real-life male (left) and female faces (right) as a function of the orthogonal\textsubscript{[comp ∥ attr]} model manipulation.

Separate groups of participants ($n = 125$ each face gender) rated male and female faces ($n = 10$ each) varied by the orthogonal\textsubscript{[comp ∥ attr]} model (Appendix U). The experimental procedure and analyses were identical with Study 7. See Appendix S for a detailed description of the methods of the study, including how I created the orthogonal\textsubscript{[comp ∥ attr]} model. For each impression, linear (gray line) and quadratic (black line) models were fitted for the ratings averaged across faces. $b_1$ and $b_2$ are unstandardized coefficients of the models. The face model represents facial information that contributes to the variance in competence impressions but not in attractiveness. Error bars denote ±SE.