Can Cinderella Become Snow White? The Influence of Perceived Trust on the Representation of Others' Faces (Postprint)

Authors: Li Qinggong, Fang Wei, Hu Chao, Shi Dejun, Hu Xiaoqing, Fu Genyue, Wang Qiandong

Date: 2023-10-09T00:00:00+00:00

Abstract

This study investigates whether perception of others' trustworthiness influences facial representation of those individuals and the underlying mechanisms. Experiment 1 had participants form impressions of target individuals as trustworthy or untrustworthy. Subsequently, reverse correlation image classification technique was used to visualize participants' mental representations of the target individuals' faces. The results showed that, regardless of whether the target individuals were male or female, high-trustworthiness targets were associated with facial representations possessing more attractive and positive traits. Experiment 2 visualized the features of facial representations of trustworthy and untrustworthy groups from a new group of participants, and conducted similarity analysis with the features of target individuals' facial representations obtained in Experiment 1. It was found that facial representation features of target individuals described as trustworthy (or untrustworthy) showed greater similarity to those of trustworthy (or untrustworthy) groups, suggesting that when people learn that others are trustworthy (or untrustworthy), they superimpose corresponding schematic features from their mind onto the physical facial features of that individual, thereby reshaping the facial representation. This study demonstrates that top-down processing plays an important role in the formation of facial representations.

Full Text

Can Cinderella Become Snow White? The Influence of Perceived Trustworthiness on Mental Representations of Faces

LI Qinggong[1], FANG Wei[2],[3], HU Chao[4], SHI Dejun[5], HU Xiaoqing[6], FU Genyue[7], WANG Qiandong[8]

(1 Zhejiang Philosophy and Social Science Laboratory for the Mental Health and Crisis Intervention of Children and Adolescents, College of Psychology, Zhejiang Normal University, Jinhua 321004, China)
(2 Department of Psychology, Zhejiang Sci-Tech University, Hangzhou 310018, China)
(3 Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton L8S 4L8, Canada)
(4 Department of Medical Humanities, School of Humanities, Southeast University, Nanjing 211189, China)
(5 School of Psychological and Cognitive Sciences, Peking University, Beijing 100871, China)
(6 The State Key Laboratory of Brain and Cognitive Sciences, Department of Psychology, The University of Hong Kong, Hong Kong 999077, China)
(7 Zhejiang Philosophy and Social Science Laboratory for Research in Early Development and Childcare, Zhejiang Key Laboratory for Research in Assessment of Cognitive Impairments, Department of Psychology, Hangzhou Normal University, Hangzhou 311121, China)
(8 Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University, Beijing 100875, China)

Abstract

People often infer others' social traits, such as trustworthiness, from a glance at their face. Whereas previous studies have focused on how different facial cues influence social perception, the present study examined whether perception of a person's trustworthiness could influence mental representations of that person's face, as well as the mechanisms underlying this process.

Two experiments were conducted. Experiment 1 was designed to test whether a target person described as trustworthy would be represented in the perceiver's mind as more attractive than the same person described as untrustworthy. Participants were instructed to form an impression about a target person's trustworthiness by viewing the person's face paired with a description labeling them as trustworthy or untrustworthy. The reverse correlation image classification (RCIC) technique was then used to visualize the participants' mental representations of the target person's face. A separate group of participants was recruited to evaluate the attractiveness and other traits of the generated mental representation images. Experiment 2 aimed to determine a possible underlying mechanism by exploring whether the mental representations of the trustworthy (or untrustworthy) target persons' faces in Experiment 1 shared more similarities with those of the trustworthy (or untrustworthy) faces at a group level (i.e., prototypes of trustworthy or untrustworthy faces). To achieve this goal, we recruited participants to complete an alternate RCIC task in which they selected which of two faces appeared more trustworthy, producing mental representation images for trustworthy and untrustworthy faces at a group level. The features

of these prototypical trustworthy and untrustworthy faces were then compared with those of the target person from Experiment 1.

In Experiment 1, mental representations of a face described as trustworthy were found to be more attractive than those of the same face described as untrustworthy. Furthermore, raters attributed additional desirable traits, such as friendly, intelligent, and positive, to the representation of the trustworthy person. In Experiment 2, we found that the mental representation of the face labeled as trustworthy in Experiment 1 shared more similarities with the prototypical trustworthy face produced in Experiment 2 than with the prototypical untrustworthy face.

In sum, our findings suggest that the perception of a person's trustworthiness can influence mental representations of that person's face. When people perceive an individual as trustworthy (or untrustworthy), they may superimpose the corresponding schema features in their minds onto the physical characteristics of the perceived individual's face, leading to a reconfiguration of the face representation. Our study underscores the importance of top-down factors in shaping face representations.

Keywords: person perception, reverse correlation image classification technology, mental representation, attractiveness, trustworthiness

People often infer a person's traits from facial appearance (Bonnefon et al., 2017; Lin et al., 2021; Sutherland & Young, 2022; Todorov et al., 2015), such as judging others' trustworthiness and competence (Todorov et al., 2009) and making positive (or negative) social evaluations of individuals (Bascandziev & Harris, 2014; Chen et al., 2014). Research has found that people generally believe that physically attractive individuals possess more positive qualities, leading to a "what is beautiful is good" stereotype (Dion et al., 1972). Additionally, knowledge about an individual's character can also change people's perception of that individual's facial attractiveness to some extent (Bliss-Moreau et al., 2008; Todorov & Uleman, 2002, 2003, 2004). A typical case is "beauty is in the eye of the beholder," where individuals tend to make more favorable evaluations of faces they like and consider honest people more physically attractive than dishonest people (Paunonen, 2006). However, whether perception of individual traits influences the mental representation of that person's face, and if so, what the underlying mechanism might be, remains unclear.

Mental representation refers to the internal reproduction of external objects in psychological activities. It reflects objective reality on the one hand, while serving as the object for further psychological processing on the other. Because mental representations can become objects of further psychological processing, mental representations of individual faces may change as attitudes toward those individuals change. According to the notion that "seeing is believing," face representations reflect a "bottom-up" visual processing mode, suggesting that people's representations of individuals' faces should only reflect the physical characteristics of those faces and should not be influenced by "top-down" processing such as impressions. However, research has shown that people's mental representations of social groups can be influenced by "top-down" stereotypes (Dotsch et al., 2008; Lloyd et al., 2020). For example, highly prejudiced participants represent Moroccan faces as more criminal and less trustworthy (Dotsch et al., 2008).

The present study investigates whether mental representations of faces differ when people form different impressions of individuals. Previous studies have confirmed that trustworthiness and attractiveness are highly correlated when both features are evaluated simultaneously (Dion, 1972; Langlois et al., 2000; Mende-Siedlecki et al., 2013). At the same time, perceptions of attractiveness and trustworthiness are important components of social perception and social interaction.

Generally, mental face representations are relatively abstract, and the rich facial features may be difficult to introspect or self-report consciously. The reverse correlation image classification (RCIC) technique can visualize the content of mental representations intuitively, revealing people's internal representations and decision-making strategies (Dotsch & Todorov, 2012; Gosselin & Schyns, 2003). This data-driven technique requires participants to choose between two noisy faces (the same base face superimposed with different random visual noise) across multiple trials (often at least 300 trials) to select the face that best represents the target individual or a specific social group. By making these selections across many trials, participants essentially provide information about the image features relevant to their mental representations. The noise patterns from the selected faces are averaged and combined with the base face to obtain the participant's classification image. Since this image displays the stimulus features that drive the social judgment of interest, it is considered an internal mental representation. In the present study, the operational definition of mental face representation is the classification image obtained using RCIC technology as a visualized mental face representation.

RCIC technology has been widely applied in social perception research to investigate how members of different social categories are represented in perceivers' minds. For example, researchers have explored how people represent outgroup members (Dotsch et al., 2008; Dotsch et al., 2011), romantic partners (Karremans et al., 2011), political candidates (Young et al., 2014), police officers (Lloyd et al., 2020), and even self-faces (Maister et al., 2021; Moon et al., 2020). Using this technique, researchers have obtained participants' mental representations of trustworthy and untrustworthy groups. Subsequently, another group of participants evaluated these two face representations and found that, compared to the representation of the untrustworthy group, participants rated the representation of the trustworthy group as more trustworthy (Dotsch & Todorov, 2012).

Building on Dotsch and Todorov (2012), the present study uses RCIC technology to examine the effect of manipulating target persons' trustworthiness on their face representations during impression formation tasks. We hypothesized that when a target person is described as trustworthy, participants' mental representations of that target person's face would be more attractive. If differences emerge in the represented faces between the two trustworthiness conditions, we can conclude that top-down trait perception influences mental representations of others' faces (participants in both conditions saw identical faces, eliminating the influence of bottom-up facial physical features). This study particularly focuses on the relationship between facial attractiveness and trustworthiness perception because previous research has confirmed a high correlation between the two features when evaluated simultaneously (Dion, 1972; Langlois et al., 2000; Mende-Siedlecki et al., 2013). At the same time, perceptions of attractiveness and trustworthiness are important components of social perception and social interaction.

2. Experiment 1: The Influence of Perceived Trustworthiness on Mental Representations of Others' Faces

Experiment 1 investigated whether mental representations of female and male target individuals' faces would be affected by the target individuals' trustworthiness. The experiment consisted of two phases. In the first phase, participants completed an impression formation task to form an impression of a target person as trustworthy or untrustworthy. Subsequently, participants completed a reverse correlation image classification task in which they needed to recall and select, from a series of two noisy faces, the face that most resembled the target person they had seen, generating face classification images (CIs) to visualize their mental representation of the target individual's face. The second phase recruited another group of participants to evaluate the traits of the classification images generated in the first phase, focusing primarily on whether the classification images produced for persons described as trustworthy would be rated as more attractive.

2.1.1 Participants

In the first phase, we recruited a total of 155 participants who were randomly assigned to four groups (2 target face genders: male and female × 2 trustworthiness levels: trustworthy and untrustworthy). The male target face trustworthy group had 40 participants (20 female participants) with a mean age of 20.38 years (SD = 1.85). The male target face untrustworthy group had 40 participants (20 female participants) with a mean age of 19.68 years (SD = 1.82). The female target face trustworthy group had 37 participants (20 female participants) with a mean age of 19.86 years (SD = 1.60). The female target face untrustworthy group had 38 participants (21 female participants) with a mean age of 20.42 years (SD = 1.95).

In the second phase, there were two tasks, and all participants recruited for this phase had not participated in the first phase. Task 1 required 40 participants (20 females, mean age = 19.98 years, SD = 1.29) to evaluate the attractiveness of the classification images generated in the first phase. Task 2 replicated and expanded the results of Task 1, recruiting 50 new participants (27 females, mean age = 21.32 years, SD = 2.51) to evaluate not only facial attractiveness but also multiple other traits (including trustworthiness, intelligence, friendliness, positive expression, meanness, greed, aggressiveness, and dominance).

All participants were university students with normal or corrected-to-normal vision. The study was approved by the Ethics Committee of Zhejiang Normal University. Participants signed written informed consent before participating and received monetary compensation after the experiment.

2.1.2 Materials

Target Person Faces. Forty participants (20 females) who did not participate in any formal experimental tasks viewed 80 face images and rated their attractiveness (1 = not at all attractive, 9 = very attractive). From these, we selected four faces with moderate attractiveness as target person faces (one male and one female) and distractor person faces (one male and one female). The average attractiveness ratings for the male target, female target, male distractor, and female distractor faces were 5.03, 5.18, 5.13, and 4.90, respectively. One-sample t-tests comparing these ratings to the midpoint of 5 yielded $t(39) = 0.11$, $p = 0.910$; $t(39) = 0.76$, $p = 0.455$; $t(39) = 0.47$, $p = 0.644$; and $t(39) = -0.41$, $p = 0.682$, respectively. The target person faces were the faces that participants needed to observe and later recognize (see experimental procedure). We selected faces with moderate attractiveness to avoid ceiling or floor effects.

Face Base Images. We morphed the target person face with a same-gender distractor person face to generate faces possessing 50% of the physical features of the target and 50% of the distractor (one male and one female face), as shown in [Figure 1: see original paper].

Stimuli for the Reverse Correlation Image Classification Task. The stimuli were generated by superimposing positive (or negative) random visual noise onto the face base images. The noise consisted of random visual patterns composed of 6 different orientations (0°, 30°, 60°, 90°, 120°, and 150°) × 5 spatial frequencies (1, 2, 4, 8, and 16 cycles/image) × 2 phases (0 and $\pi/2$) × random amplitudes (Dotsch et al., 2008). A total of 640 pairs of face stimuli were generated, with each pair containing two face stimuli superimposed with opposite noise patterns [Figure 1: see original paper]. The 640 pairs of stimuli used in this study far exceeded those used in previous similar experimental paradigms (e.g., Dotsch & Todorov, 2012). Using opposite random noise patterns also maximized the differences between the two stimuli presented in each trial, enhancing visual contrast effects and thereby reducing unnecessary experimental trials (Dotsch & Todorov, 2012). Each image was 512 × 512 pixels.

Descriptions of Trustworthy Target Persons. "Xiao Li, female (male), 20 years old, university student. Her (his) roommate says she (he) is very trustworthy and always asks for permission before borrowing others' belongings.

Once, Xiao Li found a wallet containing a large amount of cash and returned it to the owner."

Descriptions of Untrustworthy Target Persons. "Xiao Li, female (male), 20 years old, university student. Her (his) roommate says she (he) is very untrustworthy and never asks for permission before borrowing others' belongings. Once, Xiao Li found a wallet containing a large amount of cash and did not return it to the owner."

2.1.3 Experimental Procedure and Data Processing

[Figure 2: see original paper] shows the specific procedure for Experiment 1.

(1) Phase 1: Obtaining Classification Images. Participants were instructed to observe a male or female target person's face and read a brief description of that person for 3 minutes. The target person was described as trustworthy in the trustworthy condition and untrustworthy in the untrustworthy condition (see materials for descriptions of trustworthy/untrustworthy target persons). Participants were told to observe the face and read the text carefully and that there would be some related tests later. After completing a 2-minute distractor task (counting backward from 999 and writing the results on paper), participants were asked to rate the target person's trustworthiness (1 = very untrustworthy, 9 = very trustworthy) to ensure the trustworthiness manipulation was effective. Additionally, to ensure that the trust manipulation did not affect traits unrelated to trustworthiness, we also had participants rate the target person's activeness (1 = very inactive, 9 = very active). Activeness was selected as a control variable and could also test for unique psychological effects caused by the other independent variable (target face gender): we expected the general public to perceive males as more active than females.

After completing another 3-minute distractor task, participants entered the reverse correlation image classification phase to visualize their mental representation of the target person's face (Dotsch & Todorov, 2012). In each trial, two blurry faces (base face superimposed with opposite random noise patterns) were presented side by side [FIGURE:1 and FIGURE:2]. Participants needed to select the face that most resembled the target face they had seen in the first phase (the impression formation phase). There were a total of 640 trials, with randomly generated noise patterns for each trial.

The noise patterns from the 640 images selected by each participant were averaged to obtain each participant's noise pattern. These noise patterns were then averaged across all participants in each experimental condition and superimposed onto the original face base image, producing the classification images (i.e., face representations) for female and male target faces under trustworthy and untrustworthy conditions shown in [Figure 3: see original paper].

(2) Phase 2: Evaluating Classification Images. Phase 2 Task 1 aimed to examine whether individuals described as trustworthy would be associated with more attractive face representations. This task included two trials (male and female stimuli) with random presentation order. In each trial, two classification images of the same gender were presented side by side: one from the trustworthy group and one from the untrustworthy group, with random left-right positioning. Participants needed to select the more attractive face.

Task 2 replicated and expanded the results of Task 1. Participants not only evaluated facial attractiveness but also evaluated multiple other traits (including trustworthiness, intelligence, friendliness, positive expression, meanness, greed, aggressiveness, and dominance). In each trial, a pair of faces and an evaluation dimension question (e.g., "Which of the two faces below looks more trustworthy?") were presented simultaneously, and participants made their selection by pressing a key. This task had a total of 18 trials (2 genders $\times$ 9 trait evaluations), with random trial order and random left-right positioning of faces in each trial.

To explore which facial regions participants primarily used for face recognition (i.e., diagnostic regions for face recognition), we used the stat4CI toolbox developed in Matlab to conduct pixel cluster tests on the noise patterns of the four classification images described above (Chauvin et al., 2005). Following the testing method of Dotsch and Todorov (2012), we first smoothed the pixel values of the noise patterns using a Gaussian filter (standard deviation of the Gaussian = 4 pixels). Second, we performed Z-transformation on the pixel values of the smoothed noise patterns within the face region. Finally, we conducted two-tailed cluster tests on the Z-transformed data (Zcrit $\geq |2.3|$, $p < 0.05$) to reveal significant pixel clusters.

2.2 Results

2.2.1 Effectiveness of the Trustworthiness Manipulation We conducted a 2 (trustworthiness level: trustworthy vs. untrustworthy) $\times$ 2 (target face gender: male vs. female) between-subjects ANOVA on trustworthiness and activeness ratings separately. For trustworthiness ratings, only the main effect of trustworthiness level was significant, $F(1, 151) = 452.64$, $p < 0.001$, $^2 = 0.75$. Participants rated the target person in the trustworthy condition ($M = 7.30$, $SD = 1.81$) as more trustworthy than in the untrustworthy condition ($M = 1.82$, $SD = 1.34$). For activeness ratings, only the main effect of target face gender was significant, $F(1, 151) = 6.70$, $p = 0.011$, $^2 = 0.042$. Participants perceived male target persons ($M = 4.25$, $SD = 1.63$) as more active than female target persons ($M = 3.60$, $SD = 1.46$). The trustworthiness manipulation affected trustworthiness ratings but not unrelated activeness ratings, confirming the effectiveness of the manipulation.

2.2.2 Attractiveness Evaluation The classification images for trustworthy and untrustworthy conditions for female and male target faces are shown in [Figure 3: see original paper]. Regardless of whether the classification images were female or male, all participants selected the classification images from the trustworthy condition as more attractive. Chi-square tests showed $^2(1) = 40$, $p < 0.001$.

2.2.3 Multi-dimensional Trait Evaluation lists the number and proportion of raters (out of 50) who selected the classification images from the trustworthy condition for different traits, along with chi-square test results. Chi-square tests for female and male faces separately showed that more raters selected the classification images from the trustworthy condition as having more positive characteristics (attractiveness, intelligence, trustworthiness, positive expression, and friendliness). Conversely, more raters selected the classification images from the untrustworthy condition as having more negative characteristics (meanness, greed, aggressiveness, and dominance). We further included gender in the chi-square tests (when sample size in any group was less than 5, significance level was based on Fisher's exact test) and found that none of the traits were significant (see last column of), indicating that face gender did not affect participants' evaluations.

The results of the pixel cluster tests are shown in [Figure 4: see original paper]. The red and green regions represent significant areas, i.e., diagnostic regions for face recognition. These regions include the eyes, nose, mouth, and some hair, cheeks, and ears. Green clusters indicate that when the pixel values of noise in that region were smaller (making that region of the classification image darker), participants were more likely to select it as the target face. Conversely, red clusters indicate that when the pixel values of noise in that region were larger (making that region of the classification image brighter), participants were more likely to select it as the target face.

3. Experiment 2: The Mechanism of How Perceived Trustworthiness Influences Mental Face Representations

Experiment 2 examined whether the face representation features of individuals described as trustworthy (or untrustworthy) in Experiment 1 would share more similarities with the face representation features of trustworthy (or untrustworthy) groups. To generate face representations for trustworthy/untrustworthy groups (or prototypes of trustworthy/untrustworthy group faces), participants in Experiment 2 did not view individual face photos or need to recognize individual faces; they simply selected trustworthy faces from noisy faces.

3.1.1 Participants

Twenty university student participants (10 females) were recruited with a mean age of 19.95 years (SD = 1.10). All had normal or corrected-to-normal vision. The study was approved by the Ethics Committee of Zhejiang Normal University. Participants signed written informed consent before participating and received monetary compensation after the experiment.

3.1.2 Materials

The stimuli for the reverse correlation image classification task were the same as in Experiment 1.

3.1.3 Experimental Procedure

Participants completed two reverse correlation image classification tasks (male and female face tasks, with task order counterbalanced across participants) to obtain classification images for trustworthy and untrustworthy group faces (Dotsch & Todorov, 2012). Unlike Experiment 1, participants did not perform an impression formation task; they only completed the reverse correlation image classification task: selecting from two blurry faces the one that most resembled their mental image of a trustworthy face, completing a total of 640 trials. According to Dotsch and Todorov (2012), the noise from selected images averaged and superimposed onto the face base image produced the trustworthy group face representation, while the noise from unselected images averaged and superimposed onto the face base image produced the untrustworthy group face representation.

3.1.4 Data Processing

To explore the similarity between different face representation features, we extracted the pixel brightness values of the noise patterns in the face region of the classification images from both experiments, vectorized them, and calculated Pearson correlations (Dotsch & Todorov, 2012). We did not directly correlate the pixel values of the classification images themselves to avoid redundant correlations caused by the face base images, as all same-gender classification images were generated by superimposing respective noise patterns onto the same face base image. Additionally, since superimposing noise patterns can change the appearance of the face base image, the noise pattern can be considered a unique feature of different face representations.

To test the significance of correlation coefficients, we used bootstrap methods to obtain 95% confidence intervals for the correlation coefficients. If the 95% confidence interval did not contain 0, the correlation coefficient was considered significantly different from 0.

To compare differences between two correlation coefficients, we used two methods. The first adopted the method proposed by Zou (2007), using the cocor package in R software (http://www.R-project.org) (Diedenhofen & Musch, 2015) to calculate confidence intervals for differences between two correlation coefficients. If the 95% confidence interval did not include 0, the two correlation coefficients were considered significantly different. The second method used permutation tests, comparing the real difference value between correlation coefficients with a null distribution generated by random permutation. To generate the null distribution, we randomly shuffled the spatial positions of noise pixel values within the face region (e.g., swapping pixel values at positions 100 and 200) to destroy the original image structure and calculated the difference in correlation coefficients. This random permutation was performed 1000 times, generating 1000 difference values to form the null distribution. Finally, we calculated the position of the real difference value in the null distribution to compute the p-value. For two-tailed tests, values in the bottom 2.5% or top 97.5% of the null distribution were considered significant. This study focused on whether the noise pattern features of target individual face classification images in the trustworthy condition of Experiment 1 would share more similarities with the noise pattern features of trustworthy group face classification images, and whether the noise pattern features of target individual face classification images in the untrustworthy condition would share more similarities with the noise pattern features of untrustworthy group face classification images.

To further control for systematic bias caused by the same face base images, we used partial correlation methods, incorporating the vectorized pixel values of the face base images as control variables, and recalculated correlations between noise pattern features of classification images. Differences between two correlation coefficients were tested using the permutation test method described above.

Since vectorizing image pixel values loses the spatial relationships between pixel blocks, calculating Pearson correlations of vectorized pixel values may not be the best method to characterize similarity between images. We further used the ssim function in Matlab software to calculate the structural similarity index measure (SSIM) between different noise patterns to measure similarity between them (Wang et al., 2004). To test whether differences between two SSIM values reached statistical significance, we again used permutation tests, comparing the real SSIM difference value with a null distribution generated by random permutation to calculate p-values.

[Figure 5: see original paper] shows the classification images for trustworthy and untrustworthy group faces.

shows the correlation coefficients between noise patterns of different classification images and their corresponding 95% confidence intervals. Comparison of correlation coefficient differences revealed that for both male and female faces, the similarity between the noise pattern of target face classification images in the trustworthy condition and the noise pattern of trustworthy group face classification images was higher than the similarity with the noise pattern of untrustworthy group face classification images. Additionally, for male faces, the similarity between the noise pattern of target face classification images in the untrustworthy condition and the noise pattern of untrustworthy group face classification images was higher than the similarity with the noise pattern of trustworthy group face classification images. However, for female faces, the similarity between the noise pattern of target face classification images in the untrustworthy condition and the noise pattern of trustworthy group face classification images was higher than the similarity with the noise pattern of untrustworthy group face classification images.

When face base images were included as control variables, the partial correlation results between noise patterns of different classification images were consistent with the correlation results reported above, as shown in .

shows the SSIM values between noise patterns of different classification images. Comparison of differences revealed that for both male and female faces, the structural similarity between the noise pattern of target face classification images in the trustworthy condition and the noise pattern of trustworthy group face classification images was higher than the similarity with the noise pattern of untrustworthy group face classification images. Additionally, the structural similarity between the noise pattern of target face classification images in the untrustworthy condition and the noise pattern of untrustworthy group face classification images was higher than the similarity with the noise pattern of trustworthy group face classification images.

4. General Discussion

The present study used reverse correlation image classification technology to investigate the influence of perceived trustworthiness on mental representations of individual faces and its underlying mechanisms. Experiment 1 found that when the same target face was assigned different trustworthiness descriptions, the face representation of the person described as trustworthy (relative to untrustworthy) was also rated as more attractive. This result suggests that trait inference influences the perception of facial attractiveness not only through self-reported ratings but also through perceptual processes such as face representation formation. Previous studies have found that representations of social groups (e.g., male groups) are influenced not only by bottom-up visual features (e.g., larger chins) but also by top-down stereotypes (e.g., aggressiveness) (Bagnis et al., 2019; Freeman & Ambady, 2011; Freeman & Johnson, 2016; Hehman et al., 2014). Our results are consistent with previous research, demonstrating that "the face people represent in their minds is not what they actually see" and that top-down factors (such as knowledge of personality traits) can also influence how people represent others' faces.

Experiment 1 also showed that people not only perceived trustworthy individuals' face representations as more attractive but also described them as having more other positive traits. Previous research has confirmed a close connection between trustworthiness and attractiveness, indicating that perceptions of others' trustworthiness depend on their facial attractiveness (Gutierrez-Garcia et al., 2019; Oosterhof & Todorov, 2008; Willis & Todorov, 2006; Xu et al., 2012). Neuroimaging studies have also shown that judgments of facial trustworthiness and attractiveness involve similar brain regions and that this perception is spontaneous (Bzdok et al., 2011; Engell et al., 2007). Our results are consistent with previous studies. Furthermore, trustworthiness is a multidimensional trait that includes aspects such as honesty, reliability, and benevolence (Mayer et al., 1995). The present study further demonstrates that impressions of trustworthiness also influence perceptions of other positive characteristics in individuals (such as friendliness, intelligence, etc.).

Experiment 1 also explored the diagnostic regions used in face recognition, which mainly included the eyes, nose, and mouth, as well as some hair and facial contours (cheeks and ear edges) [Figure 4: see original paper]. This result is highly consistent with Dotsch and Todorov (2012), although they had participants judge faces for trustworthiness and dominance. These results suggest that when making different social judgments, people tend to extract information from similar facial regions, particularly relying on key facial regions such as the eyes, nose, and mouth. Additionally, the identification of diagnostic regions highlights that RCIC technology can provide very rich data.

Building on Experiment 1, Experiment 2 explored the potential mechanism by which perceived trustworthiness influences mental face representations. We hypothesized that when forming mental representations of target faces, individuals integrate stereotypes about a social group into the target individual. Specifically, in this study, descriptions of trustworthiness (or untrustworthiness) would lead individuals to categorize the target person into a trustworthy (or untrustworthy) group. Then, people would incorporate the stereotypical facial features of trustworthy (or untrustworthy) groups from their minds into the target person's facial features. Through this process, perceived trustworthiness levels cause perceivers' representations of the target person's objective facial physical features to become biased. The results of Experiment 2 support this hypothesis: structural similarity indices revealed that regardless of whether the target person was male or female, the face representation features of trustworthy (or untrustworthy) target persons shared more similarities with the face representation features of trustworthy (or untrustworthy) groups. However, Pearson correlation/partial correlation results showed that when the target person was female, the face representation features of untrustworthy target persons also shared more similarities with the face representation features of trustworthy groups.

This may be because calculating correlations after vectorizing image pixel values loses the spatial relationships between pixel blocks, making it unable to adequately characterize similarity between images. Additionally, people find it difficult to uglify a person. The fact that Tables 2 and 3 reveal that the similarity between untrustworthy target persons and untrustworthy groups is smaller than the similarity between trustworthy target persons and trustworthy groups confirms this point. People also tend to believe that females are more trustworthy than males (Buchan et al., 2008; Dong et al., 2018). Therefore, during the representation process, people may also add more trustworthy features to untrustworthy female target faces, which may be another reason for the unstable results for female faces. Future research needs to further explore these possibilities.

In summary, our study is the first to find that perceived trustworthiness influences mental representations of individual faces, with a trustworthy individual's face being endowed with more positive features. Therefore, Cinderella can become Snow White: even if a person is average-looking, as long as they have a good heart, the image of that person in people's minds will become more attractive.

Future research could investigate whether formed face representations affect explicit behavior. For example, do face representations formed during initial face-to-face social interactions influence subsequent non-face-to-face social interactions (such as phone calls or emails) and decisions made in business and politics? Additionally, research could explore the possibility of reshaping mental representations of individual faces. Many studies have shown that memory can be reconstructed (Barrouillet & Camos, 2014; Loftus, 1975; Mecklinger et al., 2016). As a form of memory, mental face representations should also be reconstructable. Future research could explore this reconstructive possibility and whether changes in mental face representations also alter explicit behavior or attitudes toward the target person. From a broader perspective, subsequent research could consider developing effective intervention methods to reconstruct mental representations of other-race faces and examine whether reconstructing face representations can effectively promote positive interracial interactions.

This study also has some limitations worth exploring in future research. First, face classification images obtained using RCIC technology can only be considered approximations of real mental representations because the choice of face base images and noise patterns affects the appearance of face representations (Dotsch & Todorov, 2012). Undoubtedly, using a different face base image would produce different classification images from those obtained in the current study. Compared to face base images, the choice of noise patterns has much less influence on the appearance of face representations. Dotsch and Todorov (2012) used two sets of randomly generated noise patterns (300 pairs of opposite noise patterns each) and found highly consistent results. This may be because RCIC technology typically requires many trials for averaging (generally more than 300 trials in the literature; this study used 640 trials), which reduces the influence of noise pattern selection. Even with these confounding factors, we must emphasize that studies using RCIC technology often focus on relative differences between categories (e.g., whether one category of face representation is more attractive than another) rather than the absolute appearance of a certain category of face representation. This study primarily focused on differences in mental representations of individual faces caused by trustworthy/untrustworthy manipulations or differences in similarity between trustworthy/untrustworthy individuals and trustworthy/untrustworthy group face representations. Although the face base images were the same for same-gender faces and the superimposed noise patterns were also the same, the final differences were still significant, indicating that changes in mental representations mainly originated from different descriptions of the target face's trustworthiness. Nevertheless, separating confounding factors to obtain true mental representations is worth further exploration in future research.

Second, this study mainly investigated how individual face representations are influenced by trustworthy/untrustworthy descriptions, but it is impossible to study every individual. Therefore, whether our results are specific to the two individuals selected in Experiment 1 or hold true for most individuals requires further expansion and verification in future research. In particular, this study selected individuals with moderately attractive faces, so whether the results can be generalized to faces with inherently high or low attractiveness remains to be explored.

References

Bagnis, A., Celeghin, A., Mosso, C. O., & Tamietto, M. (2019). Toward an integrative science of social vision in intergroup bias. *Neuroscience and Biobehavioral Reviews*, 102, 318−326.

Barrouillet, P., & Camos, V. (2014). *Working memory: Loss and reconstruction.* Psychology Press.

Bascandziev, I., & Harris, P. L. (2014). In beauty we trust: Children prefer information from more attractive informants. *British Journal of Developmental Psychology*, 32(1), 94−99.

Bliss-Moreau, E., Barrett, L. F., & Wright, C. I. (2008). Individual differences in learning the affective value of others under minimal conditions. *Emotion*, 8(4), 479−493.

Bonnefon, J. F., Hopfensitz, A., & de Neys, W. (2017). Can we detect cooperators by looking at their face? *Current Directions in Psychological Science*, 26(3), 276−281.

Buchan, N. R., Croson, R. T. A., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the Investment Game. *Journal of Economic Behavior & Organization*, 68(3), 466−476.

Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., Laird, A., & Eickhoff, S. B. (2011). ALE meta-analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure & Function*, 215(3-4), 209−223.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision*, 5(9), 659−667.

Chen, F. F., Jing, Y., & Lee, J. M. (2014). The looks of a leader: Competent and trustworthy, but not dominant. *Journal of Experimental Social Psychology*, 51, 27−33.

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *Plos One*, 10(3), e0121945.

Dion, K., Walster, E., & Berscheid, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, 24(3), 285−290.

Dion, K. K. (1972). Physical attractiveness and evaluation of children's transgressions. *Journal of Personality and Social Psychology*, 24(2), 207−213.

Dong, Y., Liu, Y., Jia, Y., Li, Y., & Li, C. (2018). Effects of facial expression and facial gender on judgment of trustworthiness: The modulating effect of cooperative and competitive settings. *Frontiers in Psychology*, 9, e2022.

Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, 3(5), 562−571.

Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science*, 19(10), 978−980.

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*, 100(6), 999−1014.

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, 19(9), 1508−1519.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, 118(2), 247−279.

Freeman, J. B., & Johnson, K. L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences*, 20(5), 362−374.

Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, 14(5), 505−509.

Gutierrez-Garcia, A., Beltran, D., & Calvo, M. G. (2019). Facial attractiveness impressions precede trustworthiness inferences: Lower detection thresholds and faster decision latencies. *Cognition & Emotion*, 33(2), 378−385.

Hehman, E., Ingbretsen, Z. A., & Freeman, J. B. (2014). The neural basis of stereotypic impact on multiple social categorization. *Neuroimage*, 101, 704−711.

Karremans, J. C., Dotsch, R., & Corneille, O. (2011). Romantic relationship status biases memory of faces of attractive opposite-sex others: Evidence from a reverse-correlation paradigm. *Cognition*, 121(3), 422−426.

Krosch, A. R., & Amodio, D. M. (2014). Economic scarcity alters the perception of race. *Proceedings of the National Academy of Sciences of the United States of America*, 111(25), 9079−9084.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, 126(3), 390−423.

Lin, C. J., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of english trait words. *Nature Communications*, 12(1), 5168.

Lloyd, E. P., Sim, M., Smalley, E., Bernstein, M. J., & Hugenberg, K. (2020). Good cop, bad cop: Race-based differences in mental representations of police. *Personality and Social Psychology Bulletin*, 46(8), 1205−1218.

Loftus, E. F. (1975). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal*, 15(3), 188−193.

Maister, L., de Beukelaer, S., Longo, M. R., & Tsakiris, M. (2021). The self in the mind's eye: Revealing how we truly see ourselves through reverse correlation. *Psychological Science*, 32(12), 1965−1978.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709−734.

Mecklinger, A., Rosburg, T., & Johansson, M. (2016). Reconstructing the past: The late posterior negativity (LPN) in episodic memory studies. *Neuroscience and Biobehavioral Reviews*, 68, 621−638.

Mende-Siedlecki, P., Said, C. P., & Todorov, A. (2013). The social evaluation of faces: A meta-analysis of functional neuroimaging studies. *Social Cognitive and Affective Neuroscience*, 8(3), 285−299.

Moon, K., Kim, S., Kim, J., Kim, H., & Ko, Y. G. (2020). The mirror of mind: Visualizing mental representations of self through reverse correlation. *Frontiers in Psychology*, 11, 8.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(32), 11087−11092.

Paunonen, S. V. (2006). You are honest, therefore I like you and find you attractive. *Journal of Research in Personality*, 40(3), 237−249.

Ratner, K. G., Dotsch, R., Wigboldus, D. H. J., van Knippenberg, A., & Amodio, D. M. (2014). Visualizing minimal ingroup and outgroup faces: Implications for impressions, attitudes, and behavior. *Journal of Personality and Social Psychology*, 106(6), 897−911.

Sutherland, C. A. M., & Young, A. W. (2022). Understanding trait impressions from faces. *British Journal of Psychology*, 113(4), 1056−1078.

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

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, 27(6), 813−833.

Todorov, A., & Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: Evidence from a false recognition paradigm. *Journal of Personality and Social Psychology*, 83(5), 1051−1065.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology*, 39(6), 549−562.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology*, 87(4), 482−493.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600−612.

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

Xu, F., Wu, D. C., Toriyama, R., Ma, F. L., Itakura, S., & Lee, K. (2012). Similarities and differences in chinese and caucasian adults' use of facial cues for trustworthiness judgments. *Plos One*, 7(4), e34859.

Young, A. I., Ratner, K. G., & Fazio, R. H. (2014). Political attitudes bias the mental representation of a presidential candidate's face. *Psychological Science*, 25(2), 503−510.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, 12(4), 399−413.

Note: Figure translations are in progress. See original paper for figures.