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

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

Date: 2023-05-01T00:00:00+00:00

Abstract

This study investigates whether perceptions of others' trustworthiness influence the representation of that individual's facial appearance and the underlying mechanisms. Experiment 1 had participants form impressions of target individuals as trustworthy or untrustworthy. Subsequently, using the reverse correlation image classification technique, participants' mental representations of the target individuals' faces were visualized. Results revealed that regardless of whether the target individuals were male or female, high-trustworthiness targets were associated with facial representations possessing greater attractiveness 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 analyses with the features of target individuals' facial representations obtained in Experiment 1, finding that target individuals described as trustworthy (or untrustworthy) exhibited greater similarity to the facial representation features 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 person, thereby reshaping the facial representation. This study demonstrates that top-down processing plays a crucial 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

This study investigated whether perceptions of others' trustworthiness influence mental representations of their facial appearance and the underlying mechanisms. In Experiment 1, participants formed impressions of target individuals as either trustworthy or untrustworthy. Reverse correlation image classification was then used to visualize participants' mental representations of the target faces. Results showed that, regardless of target gender, trustworthy targets were associated with more attractive and positively valenced facial representations. Experiment 2 visualized features of trustworthy and untrustworthy group prototypes in a new sample and conducted similarity analyses with the representations obtained in Experiment 1. We found that faces described as trustworthy (or untrustworthy) shared greater similarity with the trustworthy (or untrustworthy) group prototypes, suggesting that when people learn someone is trustworthy (or untrustworthy), they superimpose corresponding schematic features onto that individual's facial features, thereby reshaping the facial representation. These findings demonstrate that top-down processing plays a crucial role in the formation of face representations.

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

People frequently infer traits from facial appearance, such as judging others' trustworthiness and competence, and make positive or negative social evaluations accordingly [**?**, **?**, **?**, **?**]. Research has shown that physically attractive individuals are perceived as possessing more desirable qualities, reflecting a "what

is beautiful is good" stereotype [?]. Conversely, knowledge about a person's character can alter perceptions of their facial attractiveness [?, ?, ?, ?]. A classic example is "beauty is in the eye of the beholder," where individuals tend to evaluate liked faces more positively, perceiving honest people as more physically attractive than dishonest ones [?].

However, whether trait perception influences mental representations of faces and the underlying mechanisms remain unclear. Mental representation refers to the internal reproduction of external objects in psychological activity, which both reflects objective reality and serves as an object for further cognitive processing. Because mental representations can become objects of cognitive processing, facial representations may change as attitudes toward a person evolve. According to the "seeing is believing" notion, face representations reflect a bottom-up visual processing mode and should solely reflect physical facial features without being influenced by top-down processes such as impression formation. Nevertheless, research indicates that mental representations of social groups are shaped by top-down stereotypes [?, ?]. For example, highly prejudiced individuals represent Moroccan faces as more criminal and untrustworthy [?]. The present study examines whether mental representations of individual faces differ when people form different impressions. Previous work has shown that individuals who represent African Americans with darker skin tones and more stereotypical features allocate fewer resources to them [?], demonstrating that mental face representations directly influence decision-making [?]. Investigating the formation of mental face representations is thus crucial for understanding cognitive characteristics and behavioral patterns.

This study examined whether perceived trustworthiness affects mental representations of individual faces and the underlying mechanisms. Experiment 1 asked whether the same face described as trustworthy would be represented as more attractive than when described as untrustworthy. If differences emerged between conditions despite participants viewing identical faces, this would demonstrate that top-down trait perception influences mental face representations. We focused on the relationship between attractiveness and trustworthiness because these dimensions are highly correlated [?, ?, ?] and constitute important components of social perception and interaction.

Mental face representations are abstract and difficult to introspect or report consciously. Reverse correlation image classification (RCIC) technology can visualize mental representation content and reveal internal representations and decision strategies [?, ?]. This data-driven technique requires participants to select, across many trials (typically $ $300), which of two noisy faces (created by adding random visual noise to a base face) better represents a target individual or social group. By making these selections across numerous trials, participants provide information about the image features associated with their mental representations. The noise patterns from selected faces are averaged and combined with the base face to produce classification images, which are considered internal mental representations because they reveal the stimulus features

driving social judgments. In this study, mental face representations were operationalized as classification images obtained through RCIC. This technique has been widely used in social perception research to investigate how members of different social categories are mentally represented, including out-group members [**?**, **?**], romantic partners [**?**], political candidates [**?**], police officers [**?**], and even self-faces [**?**, **?**]. Using this technique, researchers have generated mental representations of trustworthy and untrustworthy groups, with subsequent participants rating the trustworthy group representation as more trustworthy than the untrustworthy one [**?**]. Building on [**?**], the present study used RCIC to examine how manipulating a target person's trustworthiness during impression formation affects their mental face representation. We hypothesized that trustworthy targets would be represented with more attractive faces.

Experiment 2 investigated the psychological mechanism underlying how perceived trustworthiness influences mental face representations. We hypothesized that when forming mental representations of faces, individuals integrate physical facial features (bottom-up processing) with stereotypical features of the social group to which the face belongs (top-down processing). Thus, when learning someone is trustworthy, individuals may engage in secondary processing, attributing features of their trustworthy group prototype to the target face. If correct, trustworthy (or untrustworthy) target faces should share greater similarity with trustworthy (or untrustworthy) group face prototypes. Experiment 2 used the standard RCIC procedure from [**?**] to obtain trustworthy and untrustworthy group face representations for similarity analysis with Experiment 1's representations.

Experiment 1

Experiment 1 examined whether mental representations of female and male target faces were influenced by the target's perceived trustworthiness. The experiment comprised two phases. In Phase 1, participants completed an impression formation task to establish trustworthy or untrustworthy impressions of target persons. They then performed an RCIC task, recalling and selecting from pairs of noisy faces the one that best resembled the target, generating classification images (CIs) that visualized their mental representations. Phase 2 recruited a separate group of participants to evaluate the traits of these classification images, focusing on whether faces described as trustworthy would be rated as more attractive.

2.1.1 Participants

In Phase 1, 155 participants were randomly assigned to four groups (2 target face genders: male and female $\times$ 2 trustworthiness levels: trustworthy and untrustworthy). The male trustworthy group had 40 participants (20 female; mean age = 20.38 years, SD = 1.85); the male untrustworthy group had 40 participants (20 female; mean age = 19.68 years, SD = 1.82); the female trustworthy group had 37 participants (20 female; mean age = 19.86 years, SD = 1.60); and the

female untrustworthy group had 38 participants (21 female; mean age = 20.42 years, SD = 1.95).

Phase 2 included two tasks with participants who had not completed Phase 1. Task 1 involved 40 participants (20 female; mean age = 19.98 years, SD = 1.29) who rated the attractiveness of the classification images from Phase 1. Task 2 replicated and extended Task 1 with 50 new participants (27 female; mean age = 21.32 years, SD = 2.51) who rated multiple traits including attractiveness, 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 provided written informed consent and received monetary compensation.

2.1.2 Materials

Target faces. Forty participants (20 female) who did not participate in the main experiments rated 80 face images on attractiveness (1 = not at all attractive, 9 = very attractive). From these, four faces with moderate attractiveness were selected as target and distractor faces (one male and one female for each). Mean attractiveness ratings were 5.03 for male targets, 5.18 for female targets, 5.13 for male distractors, and 4.90 for female distractors. One-sample t-tests against the midpoint of 5 revealed no significant differences: $t(39) = 0.11$, $p = 0.910$; $t(39) = 0.76$, $p = 0.455$; $t(39) = 0.47$, $p = 0.644$; and $t(39) = -0.41$, $p = 0.682$, respectively. Moderately attractive faces were selected to avoid ceiling or floor effects.

Base faces. Target and same-gender distractor faces were morphed to create faces containing 50% of each person's physical features (one male and one female base face), as shown in [Figure 1: see original paper].

[Figure 1: see original paper] illustrates the target face, distractor face, and base face for female stimuli (top row). The bottom row shows base faces superimposed with positive or negative random visual noise.

RCIC task stimuli were generated by superimposing random visual noise on base faces. Noise comprised 6 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 [**?**]. A total of 640 face stimulus pairs were generated, with each pair containing opposite noise patterns [Figure 1: see original paper]. This 640-trial procedure far exceeded previous paradigms (e.g., [**?**]), and using opposite noise patterns maximized visual contrast between stimuli, reducing unnecessary trial counts [**?**]. Each image was 512 × 512 pixels.

Trustworthy target description: "Xiao Li, female (male), 20 years old, university student. Her (his) roommate says she (he) is very trustworthy and

always asks permission before borrowing things. Once, Xiao Li found a wallet containing a large sum of cash and returned it to the owner."

Untrustworthy target description: "Xiao Li, female (male), 20 years old, university student. Her (his) roommate says she (he) is very untrustworthy and never asks permission before borrowing things. Once, Xiao Li found a wallet containing a large sum of cash and did not return it to the owner."

2.1.3 Procedure and Data Processing

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

Phase 1: Obtaining classification images. Participants were instructed to observe a male or female target face for three minutes while reading a brief description. In the trustworthy condition, targets were described as trustworthy; in the untrustworthy condition, they were described as untrustworthy (see descriptions above). Participants were told to attend carefully to both face and text, as subsequent tasks would involve related tests. After a two-minute distractor task (counting backward from 999), participants rated the target's trustworthiness (1 = very untrustworthy, 9 = very trustworthy) to ensure the manipulation was effective. To verify that the trust manipulation did not affect unrelated traits, participants also rated the target's activeness (1 = very inactive, 9 = very active). Activeness served as a control variable and allowed us to test for gender-specific effects: we expected males to be perceived as more active than females.

Following another three-minute distractor task, participants completed the RCIC phase to visualize their mental representations of the target face [**?**]. Each trial presented two noisy faces side-by-side (base face superimposed with opposite noise patterns) [FIGURE:1 and FIGURE:2]. Participants selected the face that best resembled the target seen during the impression formation phase. There were 640 trials with randomly generated noise patterns for each trial.

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

[Figure 3: see original paper] shows classification images for female and male target faces in trustworthy and untrustworthy conditions (averaged across all participants).

To identify diagnostic regions used for face recognition, we conducted pixel cluster tests on the noise patterns of these four classification images using the stat4CI toolbox in Matlab [**?**]. Following [**?**], we first applied Gaussian smoothing (SD = 4 pixels) to the noise pattern pixel values. Second, we performed

Z-transformation on smoothed pixel values within face regions. Finally, two-tailed cluster tests were conducted (Zcrit $\geq |2.3|$, $p < 0.05$) to reveal significant pixel regions.

Phase 2: Evaluating classification images. Task 1 examined whether individuals described as trustworthy would be associated with more attractive face representations. This task included two trials (male and female stimuli, presented in random order). Each trial presented same-gender classification images side-by-side: one from the trustworthy condition and one from the untrustworthy condition, with left-right position randomized. Participants selected the more attractive face.

Task 2 replicated and extended Task 1 by having participants select not only the more attractive face but also the more trustworthy, intelligent, friendly, positive, mean, greedy, aggressive, and dominant face. In each trial, a face pair and evaluation question (e.g., "Which face looks more trustworthy?") were presented simultaneously, and participants made their selection. This task comprised 18 trials (2 genders $\times$ 9 trait dimensions) with randomized trial order and face positions.

2.2.1 Effectiveness of Trustworthiness Manipulation

Separate 2 (trustworthiness: trustworthy vs. untrustworthy) $\times$ 2 (target gender: male vs. female) between-subjects ANOVAs were conducted on trustworthiness and activeness ratings. For trustworthiness ratings, only the main effect of trustworthiness was significant, $F(1, 151) = 452.64$, $p < 0.001$, $^2 = 0.75$, with higher ratings in the trustworthy condition ($M = 7.30$, $SD = 1.81$) than the untrustworthy condition ($M = 1.82$, $SD = 1.34$). For activeness ratings, only the main effect of target gender was significant, $F(1, 151) = 6.70$, $p = 0.011$, $^2 = 0.042$, with male targets ($M = 4.25$, $SD = 1.63$) rated as more active than female targets ($M = 3.60$, $SD = 1.46$). The trustworthiness manipulation affected trustworthiness ratings but not unrelated activeness ratings, confirming its effectiveness.

2.2.2 Attractiveness Ratings

Classification images for trustworthy and untrustworthy conditions are shown in [Figure 3: see original paper]. For both female and male faces, all participants selected classification images from the trustworthy condition as more attractive. Chi-square tests revealed $X^2(1) = 40$, $p < 0.001$.

2.2.3 Multi-dimensional Trait Ratings

shows the number and proportion of raters selecting trustworthy-condition classification images for each trait (out of 50 participants) and chi-square test results. Separate chi-square tests for male and female faces revealed that more raters selected trustworthy-condition faces for positive traits (attractiveness, intelligence, trustworthiness, positive expression, friendliness) and untrustworthy-

condition faces for negative traits (meanness, greed, aggressiveness, dominance). Further chi-square tests incorporating gender found no significant effects for any trait (see final column of), indicating that target gender did not influence ratings.

TABLE:1 Raters selecting trustworthy-condition classification images for different traits: number (proportion) and chi-square test results (Experiment 1 and Experiment 2) with gender incorporated.

2.2.4 Diagnostic Regions for Face Recognition

Pixel cluster test results are shown in [Figure 4: see original paper]. Red and green regions indicate significant diagnostic areas for face recognition, including eyes, nose, mouth, and some hair, cheeks, and ears. Green clusters indicate that smaller noise pixel values (making that region darker) increased selection as the target face, whereas red clusters indicate that larger pixel values (making that region brighter) increased selection.

[Figure 4: see original paper] Significant clusters revealed by pixel cluster tests.

Experiment 2

Experiment 2 examined whether facial representation features of individuals described as trustworthy (or untrustworthy) in Experiment 1 would show greater similarity to trustworthy (or untrustworthy) group prototypes. To generate these group prototypes, Experiment 2 participants did not view individual face photos or perform face recognition tasks; they simply selected trustworthy faces from noisy stimuli.

3.1.1 Participants

Twenty university students (10 female; mean age = 19.95 years, SD = 1.10) with normal or corrected-to-normal vision participated. The study was approved by the Ethics Committee of Zhejiang Normal University. Participants provided written informed consent and received monetary compensation.

3.1.2 Materials

RCIC task stimuli were identical to Experiment 1.

3.1.3 Procedure

Participants completed two RCIC tasks (male and female face tasks, order counterbalanced) to obtain classification images for trustworthy and untrustworthy group prototypes [**?**]. Unlike Experiment 1, no impression formation task was performed. Participants simply selected, across 640 trials, the face that best matched their mental representation of a trustworthy face. According to [**?**], averaging noise patterns from selected images and superimposing them on base

faces produced trustworthy group representations, while averaging noise from unselected images produced untrustworthy group representations.

3.1.4 Data Processing

To examine similarity between facial representation features, we vectorized pixel intensity values from noise patterns within face regions and computed Pearson correlations [**?**]. We analyzed noise patterns rather than classification images directly to avoid redundant correlations from identical base faces within each gender. Since noise patterns alter base face appearance, they represent unique features of each mental representation. Bootstrap methods were used to obtain 95% confidence intervals for correlation coefficients; intervals excluding zero indicated significant correlations.

Two methods compared correlation differences. First, Zou's (2007) method implemented in R's cocor package [**?**] computed confidence intervals for differences between correlations; intervals excluding zero indicated significant differences. Second, permutation tests compared real correlation differences against a null distribution. To create this distribution, we randomly shuffled spatial positions of noise pixel values (e.g., swapping values at positions 100 and 200) 1,000 times, computing correlation differences each time to generate the null distribution. The p-value was determined by the position of the real difference in this distribution, with values in the bottom 2.5% or top 97.5% considered significant for two-tailed tests. Our primary focus was whether trustworthy target faces from Experiment 1 showed greater similarity to trustworthy group prototypes than to untrustworthy prototypes, and whether untrustworthy targets showed the opposite pattern.

To control for systematic bias from identical base faces, we computed partial correlations controlling for vectorized base face pixel values, using permutation tests to compare correlation differences.

Since vectorizing pixel values loses spatial information, we also computed structural similarity index measures (SSIM) between noise patterns using Matlab's ssim function [**?**]. Permutation tests assessed differences between SSIM values.

3.2 Results

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

TABLE:2 presents correlations between noise patterns of different classification images and their 95% confidence intervals. Correlation difference tests revealed that for both male and female faces, trustworthy target faces showed greater similarity to trustworthy group prototypes than to untrustworthy prototypes. For male faces, untrustworthy target faces also showed greater similarity to untrustworthy than trustworthy prototypes. However, for female faces, untrustworthy target faces showed greater similarity to trustworthy than untrustworthy

prototypes.

TABLE:3 shows partial correlation results controlling for base faces, which were consistent with the above findings.

TABLE:4 presents SSIM values between noise patterns. Difference tests revealed that for both genders, trustworthy target faces showed greater structural similarity to trustworthy than untrustworthy group prototypes. Additionally, untrustworthy target faces showed greater similarity to untrustworthy than trustworthy prototypes.

General Discussion

This study used RCIC to investigate how perceived trustworthiness influences mental representations of individual faces and the underlying mechanisms. Experiment 1 demonstrated that the same target face described as trustworthy (versus untrustworthy) was represented as more attractive, suggesting that trait inference affects not only self-reported attractiveness ratings but also perceptual processes involved in forming face representations. Previous research has shown that group face representations (e.g., male groups) are influenced by both bottom-up visual features (e.g., larger jaws) and top-down stereotypes (e.g., aggressiveness) [**?**, **?**, **?**, **?**]. Our findings align with this work, demonstrating that "what we represent in our minds is not what we actually see" and that top-down factors (e.g., knowledge of personality traits) can influence mental representations of others' faces.

Experiment 1 further showed that trustworthy individuals were represented not only as more attractive but also as possessing more other positive traits. Research has confirmed close links between trustworthiness and attractiveness, with trust judgments relying on facial attractiveness [**?**, **?**, **?**, **?**]. Neuroimaging studies indicate that trustworthiness and attractiveness judgments involve similar brain regions and occur spontaneously [**?**, **?**]. Our results are consistent with these findings. Moreover, trustworthiness is a multidimensional construct including honesty, reliability, and benevolence [**?**], and our study shows that trust impressions also influence perceptions of other positive characteristics (e.g., friendliness, intelligence).

Experiment 1 also identified diagnostic regions for face recognition, primarily including eyes, nose, mouth, and some hair and facial contours (cheeks and ear edges) [Figure 4: see original paper]. These results align with [**?**], despite their focus on trustworthiness and dominance judgments, suggesting that people extract information from similar facial regions across different social judgments, particularly relying on these key features. The identification of diagnostic regions highlights the richness of information provided by RCIC technology.

Experiment 2 explored the mechanism underlying how perceived trustworthiness influences mental face representations. We hypothesized that when forming mental representations, individuals integrate social group stereotypes into target

representations. Specifically, trustworthy (or untrustworthy) descriptions lead individuals to categorize targets into trustworthy (or untrustworthy) groups, then incorporate stereotypical facial features of those groups into the target's features, biasing representations of objective facial characteristics. Experiment 2 supported this hypothesis: SSIM analyses revealed that trustworthy (or untrustworthy) target faces shared greater similarity with corresponding group prototypes for both genders. However, Pearson/partial correlation analyses showed that for female faces, untrustworthy targets also resembled trustworthy prototypes. This discrepancy may arise because vectorizing pixel values loses spatial information, making it a suboptimal similarity measure. Additionally, people may be reluctant to uglify others, as evidenced by weaker similarity between untrustworthy targets and untrustworthy prototypes compared to trustworthy targets and trustworthy prototypes. People also perceive females as more trustworthy than males [**?**, **?**], potentially leading them to add more trustworthy features even to untrustworthy female faces, which may explain the unstable female results. Future research should explore these possibilities.

Future studies could investigate whether formed face representations influence explicit behaviors. For example, do face representations formed during initial face-to-face interactions affect subsequent non-face-to-face decisions in business or political contexts? Additionally, research could examine whether mental face representations can be reshaped, as many studies show memory can be reconstructed [**?**, **?**, **?**], and mental face representations, as a form of memory, should be similarly malleable. Future work could explore this reconstructive potential and whether reshaping representations alters explicit behaviors or attitudes toward targets. From a broader perspective, subsequent research could develop effective interventions to reconstruct mental representations of other-race faces and examine whether such reconstruction promotes positive interracial interactions.

This study has limitations warranting further exploration. First, classification images obtained through RCIC are only approximations of true mental representations, as base faces and noise selection influence representation appearance [**?**]. Different base faces would produce different classification images. However, noise selection has minimal impact compared to base faces, and [**?**] found consistent results using two sets of random noise (300 opposite-noise pairs each). RCIC typically requires many trials (usually >300; our study used 640), which reduces noise selection effects. Despite these confounds, RCIC research typically focuses on relative differences between categories (e.g., whether one category's representation is more attractive than another's) rather than absolute appearance. Our study focused on differences between trustworthy/untrustworthy conditions and similarities between individual and group representations. Although same-gender base faces and noise patterns were identical, significant differences emerged, indicating that mental representation changes primarily stemmed from different trustworthiness descriptions. Nevertheless, isolating confounding factors to obtain true mental representations remains an important direction for future research. Second, this study investigated effects of

trustworthiness descriptions on individual face representations but cannot test all possible individuals, so whether our findings generalize beyond the two individuals used in Experiment 1 requires further validation. Our selection of moderately attractive faces also raises questions about generalizability to faces with high or low attractiveness.

In summary, this study is the first to demonstrate that perceived trustworthiness influences mental representations of individual faces, with trustworthy individuals being represented with more positive features. Thus, Cinderella can become Snow White: even someone with average appearance will be perceived as more attractive if they possess a good heart.

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), 329–343.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision*, 5, 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(4), 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, 2022.

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), 378–387.

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), 1205–1220.

Loftus, E. F. (1975). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal*, 15(3), 6–10.

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.

R Core Team. (2015). R: A language and environment for statistical computing. http://www.R-project.org/

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), e009123.

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.