

Trustworthiness judgments without the halo effect: A data-driven computational modeling approach

Perception

2023, Vol. 52(8) 590–607

© The Author(s) 2023

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/03010066231178489

journals.sagepub.com/home/pec**DongWon Oh** 

National University of Singapore, Singapore

Nicole Wedel

Princeton University, USA

Brandon Labbree 

University of Michigan, USA

Alexander Todorov

University of Chicago Booth School of Business, USA

Abstract

Trustworthy-looking faces are also perceived as more attractive, but are there other meaningful cues that contribute to perceived trustworthiness? Using data-driven models, we identify these cues after removing attractiveness cues. In Experiment 1, we show that both judgments of trustworthiness and attractiveness of faces manipulated by a model of perceived trustworthiness change in the same direction. To control for the effect of attractiveness, we build two new models of perceived trustworthiness: a subtraction model, which forces the perceived attractiveness and trustworthiness to be negatively correlated (Experiment 2), and an orthogonal model, which reduces their correlation (Experiment 3). In both experiments, faces manipulated to appear more trustworthy were indeed perceived to be more trustworthy, but not more attractive. Importantly, in both experiments, these faces were also perceived as more approachable and with more positive expressions, as indicated by both judgments and machine learning algorithms. The current studies show that the visual cues used for trustworthiness and attractiveness judgments can be separated, and that apparent approachability and facial emotion are driving trustworthiness judgments and possibly general valence evaluation.

Corresponding author:

DongWon Oh, National University of Singapore, 9 Arts Link, AS4-02-17, Singapore 117572, Singapore.

Email: doh@nus.edu.sg

Keywords

social perception, trustworthiness, judgments, face perception, halo effect, attractiveness

Date Received: 7 December 2022; accepted: 9 May 2023

Trustworthiness judgments based on appearance affect a wide range of real-world outcomes. A person who is judged as trustworthy due to their appearance ‘has it easy’ in many domains. To individuals with “trustworthy” looks (as opposed to those with “untrustworthy” looks), people are more willing to loan money (Duarte et al., 2012), to pay more money for the same service (Ert et al., 2016), to give a second chance after a misconduct (Gomulya et al., 2017), and sentence leniently in court (Porter et al., 2010; Stewart, 1980, 1985; Wilson & Rule, 2015) (for review, see Todorov et al., 2015). To understand these effects and ultimately rectify these biases, it is important to understand what visual facial “cues” contribute to perceived trustworthiness.

Theoretically, this is an important question because trustworthiness judgments are also one of the best proxies for valence evaluation of faces (Oosterhof & Todorov, 2008; Todorov, 2008). As a general rule, valence evaluation (i.e., overall positive versus negative impressions of someone) is estimated as a linear combination of multiple judgments and accounts for a large amount of the variance of social judgments (Oosterhof & Todorov, 2008; Todorov & Oh, 2021). Importantly, this linear combination is highly correlated with trustworthiness judgments (>.90), even when the latter are not part of the linear combination (Oosterhof & Todorov, 2008; Todorov, 2008). As a result, trustworthiness judgments are a good starting point to study the cues driving general valence evaluation of faces. Here we are concerned with identifying cues other than attractiveness that contribute to trustworthiness judgments and potentially general valence evaluation.

It is well known that judgments of trustworthiness and attractiveness are highly correlated with each other (Todorov, 2008; Todorov et al., 2013). Moreover, there are dozens and dozens of studies summarized in meta-analyses showing that physical attractiveness puts a person in an overall positive light (Dion et al., 1972; Eagly et al., 1991). This halo effect also includes making the person appear trustworthy. The halo effect has been studied in psychology for over 100 years now (Thorndike, 1920). In the context of face evaluation, one implication of the halo effect is that any positive evaluations could be attributed to attractiveness, as long as there is a positive correlation between the latter and these evaluations. In fact, outside of the world of face evaluation research, attractiveness is often considered the most important (and occasionally the only) attribute that matters in evaluation of appearance (Olivola & Todorov, 2017). Thus, theoretically, it is important to find out whether there are other meaningful cues besides attractiveness that contribute to the evaluation of faces. Practically, to the extent that perceived trustworthiness is correlated with attractiveness, it is hard to draw causal conclusions about the role of perceived trustworthiness, because any effect of the latter on outcomes (e.g., loan decisions) may be attributed to attractiveness. Here we present new models of perceived trustworthiness that control for the effects of attractiveness.

Developments in computational modeling of judgments from faces make it possible to control for any set of cues, as long as these cues map onto specific judgments (Todorov & Oh, 2021). Because the models of judgments are vectors within the same statistical multi-dimensional space, it is possible to control for correlations between the vectors. A recent example involves judgments of perceived competence from faces (Oh et al., 2019). Much like the halo effect in trustworthiness judgments, face-based competence judgments are also positively correlated with physical attractiveness. However, using models of judgments of competence and judgments of attractiveness, it is possible to extract facial information associated with competence that is free

of the halo effect of attractiveness (Oh et al., 2019). Using these new models of competence judgments, one can make a person appear more competent but not more attractive, or less competent but not less attractive.

Utilizing this statistical modeling approach, the current study builds and validates new face models of trustworthiness judgments controlling for attractiveness. Specifically, we test whether faces manipulated by the models to appear more trustworthy are perceived as more trustworthy but not more attractive. Moreover, after removing attractiveness cues, we also identify facial cues that systematically vary with trustworthiness judgments. One candidate set of cues is emotional expressions signaling approach/avoidance behavior, consistent with the notion that trustworthiness evaluation is an extension of systems for recognition of emotions signaling approach/avoidance behavior (Todorov, 2008).

Experiment 1

In Experiment 1, we establish that facial information associated with trustworthiness judgments overlaps with facial information associated with attractiveness judgments. To test this idea, we first manipulated faces on perceived trustworthiness, using a statistical model derived from human ratings of faces on trustworthiness (Oosterhof & Todorov, 2008; Todorov et al., 2013). Participants were then asked to rate the manipulated facial images on both apparent trustworthiness and attractiveness. We expected the judgments to co-vary; faces made to appear trustworthy or untrustworthy would also appear attractive or unattractive, respectively.

Method

Participants. Participants participated in an online study for monetary compensation via Amazon Mechanical Turk. To estimate the necessary sample size, we conducted simulation-based power analysis. We used existing data collected using similar independent and dependent variables as well as face stimuli (Oh et al., 2019; data available on <https://osf.io/ygzx3>). In the power analysis, we considered the structure of our statistical model (mixed-effects model with one fixed and two random factors) and the data to be collected (data derived from a within-subjects design with 175 trials per participants; $175 = 25$ face identities * 7 manipulation levels) (see *Stimuli*, *Procedure*, and *Analysis* in what follows for details). We used R package *simr* (Green & MacLeod, 2016) which is based on *lme4*, an R Package widely used for mixed-effects modeling (Bates et al., 2015) in the R environment (R Core Team, 2018). Simulation based on the specific data structure and previously found effect sizes found that 15 or more reliable participants will afford a power of 95% to test the effect of model manipulation on impression ratings of faces. We aimed for 20 initial participants, stopping data collection at $n \geq 20$. We overshot because we expected a few participants with unreliable responses, whose data would be removed (see *Procedure* in what follows for details). Two separate groups of participants judged manipulated face images on one of the two social dimensions: trustworthiness ($n = 21$; M age = 35.67, standard deviation [SD] age = 9.75; 7 female, 14 male; 1 Asian, 3 Black, 17 White) or attractiveness ($n = 22$; M age = 39.73; SD age = 14.39; 11 female, 11 male; 2 Asian, 4 Black, 15 White, 1 other).

Stimuli. To generate faces that varied on trustworthiness judgments, we used FaceGen Modeller 3.2 (Singular Inversions). FaceGen utilizes a statistical multidimensional face space (Blanz & Vetter, 1999; Valentine, 2001). In FaceGen each face is a vector, in our case, a vector in a 100-dimensional space. Fifty numbers on 50 dimensions (i.e., parameters) determine the face shape and fifty numbers on the other 50 dimensions determine the face reflection (i.e., color and texture). These parameters had been extracted to explain a large variance across actual individual human faces by the FaceGen team. Each parameter is independent of the others (i.e., changing one

parameter does not affect any other), corresponding to a set of holistic visual changes on the face. In this framework, Todorov et al. (2013) built multiple data-driven statistical face models, each representing a specific social judgment from the face (e.g., judgments of trustworthiness). The trustworthiness-judgment model, for example, captures the changes in facial information that co-vary with changes in actual human ratings of perceived trustworthiness. The model is data-driven, because the stimuli (facial images) are randomly generated and the model is built based on the ratings of these stimuli. With this model, one can take a novel face and make it appear “more trustworthy” or “less trustworthy” (for technical details, see Oosterhof & Todorov, 2008; Todorov & Oh, 2021). We applied the standard model of trustworthiness judgments (Oosterhof & Todorov, 2008; Todorov et al., 2013) to 25 novel identities generated by FaceGen. The 25 identities were generated to appear different from each other. Each identity was then projected at -3 , -2 , -1 , 0 , 1 , 2 , and 3 *SDs* on the dimension of the standard trustworthiness model. Each *SD* represented the associated amount of change in trustworthiness ratings by human raters, relative to the average face, calculated in previous work (Todorov et al., 2013). The final stimulus set consisted of 175 face images (25 identities * 7 manipulation levels). See Figure 1 for a sample identity varying on the trustworthiness-judgment dimension. All stimuli in Experiment 1 and following experiments are publicly available on the Open Science Framework (OSF; <https://osf.io/sd4y7/>).

Procedure. In this study and all following studies, the study protocol was approved by the Institutional Review Board of Princeton University. All participants gave informed consent prior to participation. Participants provided basic demographic information (i.e., age, sex, and race) and were told that they would be asked to provide their intuitive judgments of the character of various individuals based on their image. On each trial, participants were presented with a face image and a prompt above the image, reading “How [trait] is this person?—Not at all [trait] (1) —Extremely [trait] (7)” with [trait] being either “trustworthy” or “attractive” for each participant group. In the instruction, participants were told that there was no right or wrong answer, and encouraged to rely on their gut feeling. To assess the within-rater reliability, we presented 25 extra face images randomly chosen from the 175 total face images. In total, each participant completed 200 trials (175 + 25). Each participant’s within-rater reliability was defined as the Pearson

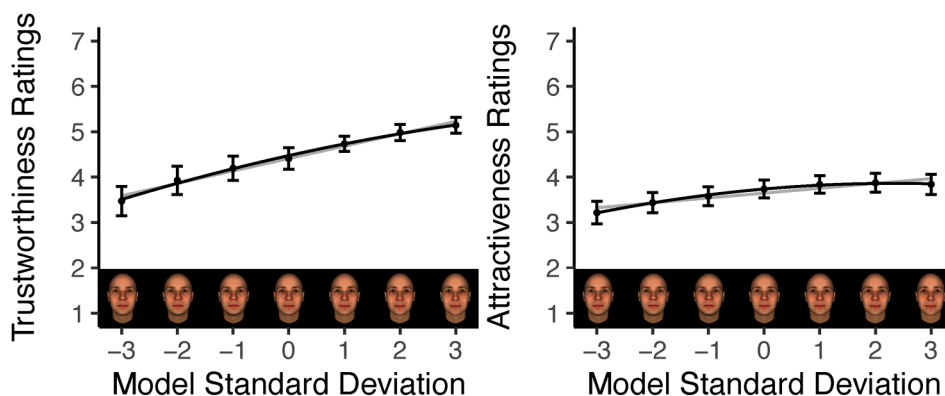


Figure 1. Judgments of trustworthiness and attractiveness of faces are manipulated by a statistical model of perceived trustworthiness (Experiment 1). When faces were manipulated to appear more trustworthy, they were perceived as both more trustworthy (left) and more attractive (right). The lines denote the linear and quadratic fit across all data points. The error bars denote the standard errors across participants. The lines and error bars are for visualization only: The actual analyses were conducted using mixed-effects models to consider the idiosyncrasies of the participants and face identities. Sample faces originating from one identity are displayed at the bottom of the graph above the model manipulation level as a reference.

correlation between ratings on the first and second presentation of the 25 repeated images. A positive correlation indicates an overall consistent response to the stimuli from that rater. Unreliable raters were defined as those with within-rater reliability less than or equal to 0. After data collection, we excluded unreliable raters from further analyses. As a result, 5 participants and 2 participants were excluded from the trustworthiness-rating and the attractiveness-rating conditions, respectively. The final samples were 16 and 20 participants, respectively. Including all participants' data did not change the results (see Supplemental Materials). Across raters, we found a high level of consensus in both trustworthiness (intraclass correlation coefficients (ICC)=0.71, Cronbach's $\alpha=0.77$) and attractiveness ratings (ICC=0.66, $\alpha=0.74$). The level of consensus was high even when all participants were considered (Supplemental Tables 1 and 2).

Analysis. We predicted trustworthiness ratings of faces using the trustworthiness-model manipulation level (ranging from -3 to $+3$ with the interval of 1) as the independent variable. We also predicted attractiveness ratings of the same faces using the trustworthiness-model manipulation level. To consider the potential idiosyncratic effect of participants and face identities, we ran mixed-effects regressions with crossed random factors (Baayen et al., 2008) for participant and face identity using *lme4* R package (Bates et al., 2015). Specifically, we treated participants and face identities as random variables and included by-participant and by-identity intercepts (rating \sim level $+(1|\text{participant})+(1|\text{face identity})$). Models of the same structure have been found suitable for considering individual variability in face evaluation (Oh, Grant-Villegas et al., 2020). Statistical significance was determined via Satterthwaite approximation using *lmerTest* R package (Kuznetsova et al., 2017). We also ran mixed-effects regressions with a quadratic term in addition to the linear term; the results were consistent with our predictions (see Supplemental Results for details). The data, as well as the code for analysis and figure-generation, for all experiments are available on OSF (<https://osf.io/sd4y7/>).

Results and Discussion

Because the models of judgments are vectors within the same multidimensional statistical space, it is straightforward to compute their similarity (Todorov et al., 2013; Todorov & Oh, 2021). In the case of the models of perceived trustworthiness and attractiveness, their correlation is $r=0.53$, indicating high redundancy in the facial cues used for these respective judgments. Consistent with this redundancy, as shown in Figure 1, faces manipulated by the standard trustworthiness model were perceived as both more trustworthy and more attractive when their manipulated trustworthiness increased. The results for trustworthiness judgments ($B=0.27$, $SE=0.01$, 95% CI [0.25, 0.30], $t=22.56$, $p<.001$; Figure 1) replicate previous validations of this model (Oosterhof & Todorov, 2008; Todorov et al., 2013). The results for the attractiveness ratings ($B=0.11$, $SE=0.01$, 95% CI [0.09, 0.13], $t=10.63$, $p<.001$) indicate that these ratings increased or decreased in the same direction as the trustworthiness ratings, though the effect of the manipulation was not as large as the effect on the latter ratings. These results suggest that under natural circumstances, facial information associated with trustworthiness judgments covaries with facial information associated with attractiveness judgments. This is consistent with studies demonstrating the halo effect of attractiveness on interpersonal trust (Dion et al., 1972; Eagly et al., 1991).

Experiment 2

Experiment 1 showed the standard model of trustworthiness could make a face appear trustworthy or untrustworthy, as well as attractive or unattractive at the same time. This suggests that facial information associated with trustworthiness judgments overlaps with facial information associated with attractiveness judgments. It remains unclear, however, whether there is other meaningful facial

information associated with trustworthiness judgments, independent of attractiveness judgments. If there is no other meaningful facial information, then removing attractiveness-associated facial information from the “trustworthiness model” will eliminate the impact of the model on people’s trustworthiness judgments. On the other hand, if there is facial information other than attractiveness that is contributing to trustworthiness judgments, then removing attractiveness-associated facial information from the “trustworthiness model” will not entirely remove the impact of the model on people’s trustworthiness judgments.

In order to test these possibilities, in Experiment 2, we modified the ‘standard’ model of trustworthiness judgments (used in Experiment 1). Specifically, the modified model did not positively covary facial information used for trustworthiness judgments with facial information used for attractiveness judgments. By subtracting the model of attractiveness judgments from the standard model of perceived trustworthiness (Todorov et al., 2013), the facial information associated with trustworthiness judgments was forced to *negatively* covary with facial information associated with attractiveness, at least in terms of the statistical models of these judgments. Using this subtraction model, we then tested whether trustworthiness ratings of manipulated faces still varied in the predicted direction, whereas the attractiveness ratings of these faces did not. That is, faces manipulated to appear more trustworthy should be judged as more trustworthy but not more attractive (and perhaps as less attractive because of the negative correlation between the subtraction model and the attractiveness model).

Further, we explored whether the remaining facial information (after removing attractiveness cues) that contributes to trustworthiness judgments is related to perceived approachability and emotion. Todorov (2008) has argued that trustworthiness judgments are an attempt to infer whether to approach or avoid a person; and empirical studies show that these judgments are highly correlated with judgments of approachability (Adolphs et al., 1998; Sutherland et al., 2013; Todorov, 2008). Todorov (2008) has also argued that trustworthiness judgments are derived from facial cues resembling emotional expressions signaling approach/avoidance behavior. Specifically, whereas perceptions of a happy expression are strongly positively correlated with trustworthiness judgments (even among “emotionally neutral” resting faces, some appear more “smily” than others), perceptions of an angry expression are negatively correlated (Oosterhof & Todorov, 2008, 2009; Said et al., 2009). We tested whether these two types of facial information—approachability cues and emotion cues—“survived” the removal of the halo effect, and were still associated with trustworthiness judgments. Specifically, faces manipulated by the subtraction model to appear more trustworthy should be perceived as more approachable and happier.

To quantitatively measure the perceived amount of happiness-related facial information, we employed a two-pronged approach utilizing both human judgments and machine learning (ML) algorithms. This approach helps mitigate any potential biases that human observers may exhibit when assessing the variation in faces with regard to perceived trustworthiness. For instance, human observers may perceive a face as trustworthy due to the conflation of positivity associated with both smiling and trustworthiness cues (whether or not the halo effect is present). In contrast, ML algorithms are immune to such biases. If lay human judgments and ML outcomes converge in the expected direction (i.e., a correlation between “facial trustworthiness” in the absence of the halo effect and the amount of estimated smiling, as measured by both human judgments and machine estimates), we can be more confident in our conclusion that smiling-related facial information indeed underlies trustworthiness judgments, in addition to the attractiveness halo, as compared to relying solely on human judgments.

Method

Participants. Participants participated in an online study for monetary compensation via Amazon Mechanical Turk. In the “trustworthiness” and “attractiveness” conditions, we aimed for three

times larger sample than in Experiment 1, stopping after $n = 60$, because the effect of the new “subtraction” model on judgments was expected to be smaller than the effect of the standard trustworthiness model. This is based on the removal of the halo effect of attractiveness, which is a substantive driver of trustworthiness judgments. In the “approachability” and “emotionality” conditions, we aimed for two times larger sample than in Experiment 1, stopping after $n = 40$, because the effect of the subtraction model on approachability and emotionality ratings was expected to be bigger than the effect on trustworthiness and attractiveness ratings. These predictions follow previous findings obtained using the “subtraction” model of competence judgments controlling for the attractiveness halo ([competence—attractiveness]) (Oh et al., 2019). Four separate groups of participants judged faces manipulated by the subtraction model on four different social dimensions: Trustworthiness ($n = 63$; M age = 37.59, SD age = 11.14; 24 female, 39 male; 4 Asian, 10 Black, 45 White, 4 other), attractiveness ($n = 61$; M age = 36.56; SD age = 11.32; 31 female, 30 male; 1 Native American, 3 Asian, 6 Black, 51 White), approachability ($n = 48$, M age = 35.48, SD age = 9.96; 12 female, 35 male, 1 nonbinary; 1 American Native, 2 Asian, 16 Black, 1 Native Hawaiian or Pacific Islander, 26 White, 1 mixed, 1 other), and emotionality ($n = 49$, M age = 38.63, SD age = 13.08; 24 female, 25 male; 5 Asian, 4 Black, 37 White, 2 mixed, 1 other).

Stimuli. To generate a new set of stimuli, we built a new model of trustworthiness judgments that are not positively correlated with attractiveness (i.e., the “subtraction model”). Specifically, we subtracted the “attractiveness model” parameters from the “trustworthiness model” parameters. Much like a face represented in the 100-dimensional FaceGen face space with 100 coordinates, each of these face models has 100 parameters, each of which represents the amount of change on that dimension (representing the amount of specific holistic changes in facial appearance) associated with a particular social judgment (e.g., trustworthiness). Because the “attractiveness model” and the “trustworthiness model” reside in the same statistical space (i.e., they have the same number of coordinates, corresponding to each other), one can subtract one model from the other. When the two models are positively correlated, this procedure creates a new model, representing face information positively associated with one judgment (in our case, perceived trustworthiness), while representing face information negatively associated with the other judgments (in our case, perceived attractiveness). To generate facial images, we used the faces of 25 different identities as we did in Experiment 1. Each identity was projected at -3 , -2 , -1 , 0 , 1 , 2 , and 3 SD on the dimension of the subtraction model ([trustworthiness—attractiveness]). The final stimulus set consisted of 175 face images (25 identities * 7 manipulation levels), as in Experiment 1. See Figure 2 for a sample identity varying on the subtraction dimension.

Procedure. The study followed the same procedure as in Experiment 1. For trustworthiness, attractiveness, and approachability ratings, participants were asked “How [trait] is this person?—Not at all [trait] (1)—Extremely [trait] (7)” with [trait] being either trustworthy, attractive, or approachable. For emotion ratings, participants were asked “Rate the expression of this person.—Extremely angry (1)—Extremely happy (7).” After data collection, as in Experiment 1, we excluded unreliable participants with within-rater reliability that is less than or equal to 0 from further analyses: 9, 8, 8, and 5 participants from the trustworthiness, attractiveness, approachability, and emotion ratings conditions, respectively. As a result, we had 54, 53, 40, and 44 final participants in each group. Including all participants’ data did not change the results (see Supplemental Materials for details). Across raters, we found a high level of consensus in all types of judgments: trustworthiness (ICC = 0.89, Cronbach’s $\alpha = 0.91$), attractiveness (ICC = 0.85, $\alpha = 0.89$), approachability (ICC = 0.89, $\alpha = 0.92$), and emotion ratings (ICC = 0.97, $\alpha = 0.97$). As in Experiment 1, the level of consensus was high even when all participants were included (Supplemental Tables 1 and 2).

Analysis. We ran mixed-effects regressions with cross-random factors of participants and face identities, predicting human judgment ratings from the model level. The regression models had

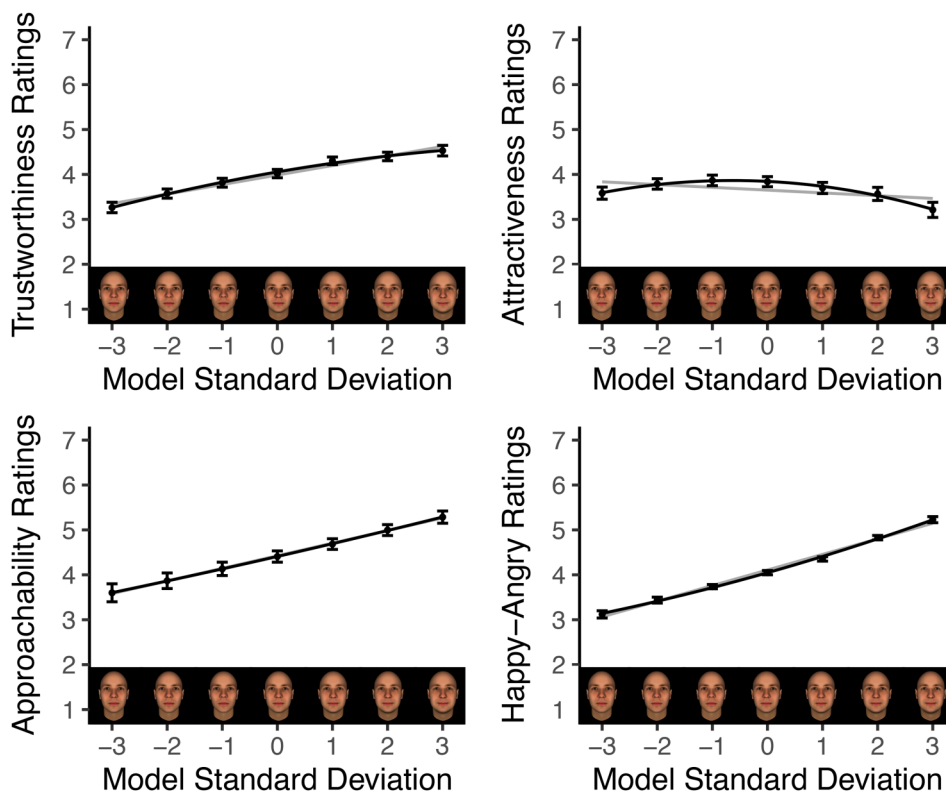


Figure 2. Judgments of trustworthiness, attractiveness, approachability, and emotional expressions of faces manipulated by a statistical model of perceived trustworthiness subtracting a model of attractiveness (Experiment 2). To remove the halo effect of attractiveness, we built a new model of trustworthiness judgments that manipulated facial information related to trustworthiness in the opposite direction to facial information related to attractiveness judgments (see main text for details). The resulting “subtraction model” of trustworthiness judgments could make faces appear more trustworthy (top left). Notably, these trustworthy-looking faces were not more attractive (top right). The trustworthy-looking faces were rated as more approachable (bottom left) and happier (bottom right). These findings suggest that facial information related to approachability and emotion was preserved in the faces, even in the absence of attractiveness cues. The lines denote the linear and quadratic fit across all data points. The error bars denote the standard errors across participants. The lines and error bars are for visualization only; the actual analyses were conducted using mixed-effects models to consider the idiosyncrasies of the participants and face identities. Sample faces originating from one identity are displayed at the bottom of the graph above the model manipulation level.

the same structure as those in Experiment 1. To quantify smile-related facial information in an objective manner, in addition to analyzing human ratings, we extracted relevant cues using an ML algorithm. We extracted facial gestural features from all face images using Py-Feat (Jolly et al., 2021), a tool for detecting and extracting various facial features from images and videos. We extracted three features from the face images. The first two are estimates of Action Units (AUs), based on the facial action coding system (Friesen & Ekman, 1978), specifically, AU6 (cheek raiser) and AU12 (lip corner puller). These gestures, when simultaneously occurring, are strongly associated with the expression of happiness or the perception thereof. The extraction algorithm is based on XGBoost (Chen & Guestrin, 2016), a visual classifier based on ensemble learning. The third feature is an estimate of emotional expression, specifically, the estimate of facial gestures of “happiness.” This algorithm is based on the residual masking network, a pretrained and tested

convolutional neural network (Pham et al., 2021). These computer algorithms have been trained on numerous face images and rigorously validated (Jolly et al., 2021; Pham et al., 2021). While our face images only had faces with resting gestures, if there was any temporally stable visual feature associated with the perceptions of happiness (e.g., high mouth corners), they should be detected by these algorithms. For example, a bigger value in an AU12 estimate would indicate that the face has an appearance that suggests a strong pull in lip corners (e.g., high mouth corners).

As in other statistical models, we predicted each of these three ML-derived estimates using the trustworthiness-model manipulation level as the independent variable. However, for these objective machine-extracted measures, we excluded the participant intercept term from the multilevel model because there was no human participant variable to consider in these models (i.e., $\text{estimate} \sim \text{level} + (1|\text{face identity})$). In addition to the analysis of Experiment 2 data, we post hoc ran the same analysis on the data from Experiment 1 (Figure 3 top) to confirm whether the variation in smile-related information was present in the original model of perceived trustworthiness.

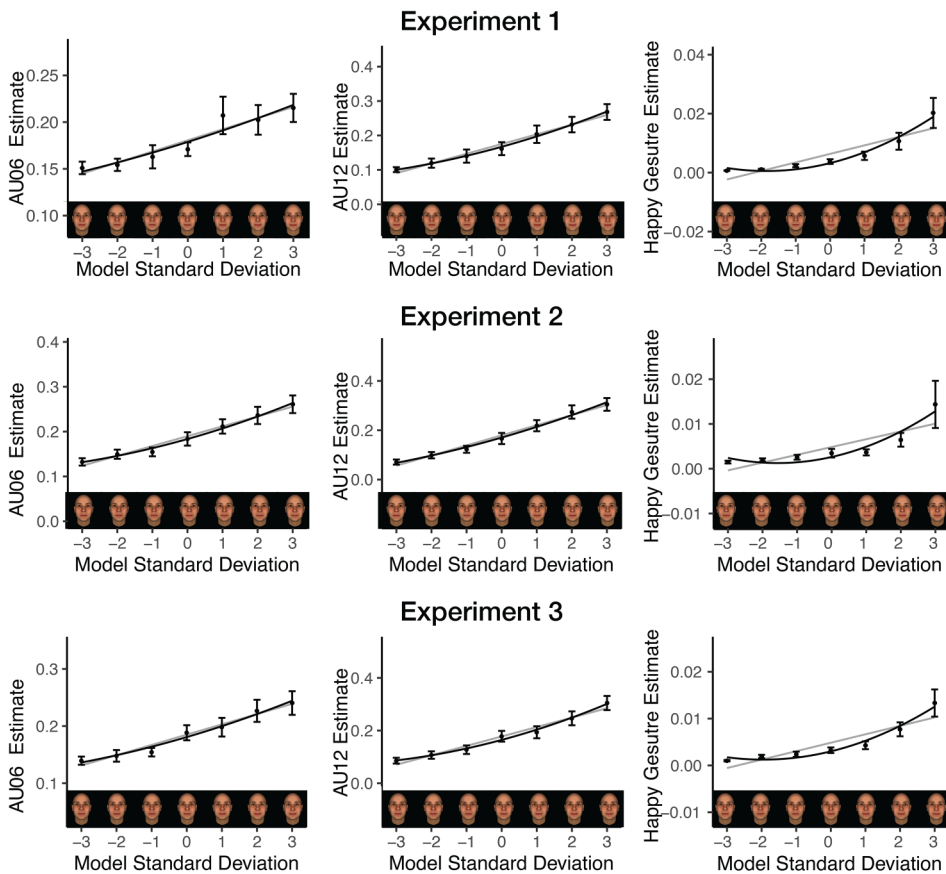


Figure 3. Machine-extracted feature estimates from faces manipulated by a statistical model of perceived trustworthiness (Experiment 1, top), a model of perceived trustworthiness subtracting a model of attractiveness (Experiment 2, middle), and a model of perceived trustworthiness orthogonal to a model of attractiveness (Experiment 3, bottom). Objectively estimated amounts of muscle group activity (estimate of AU6, cheek raiser, and AU12, lip corner puller) and emotional gesture (estimate of happiness gesture), derived from computer algorithms, show that faces manipulated to appear more trustworthy-looking appeared to be more smiling or “happier.”

Results and Discussion

The subtraction [trustworthiness—attractiveness] model could manipulate faces to appear more or less trustworthy in the expected direction ($B = 0.21$, $SE = 0.01$, 95% CI [0.2, 0.22], $t = 33.37$, $p < .001$; Figure 2). This suggests that participants used meaningful facial cues other than attractiveness to make trustworthiness judgments. In fact, attractiveness did not increase as faces were manipulated to appear more trustworthy ($B = -0.06$, $SE = 0.01$, 95% CI [-0.07, -0.05], $t = -9.14$, $p < .001$). Rather, trustworthy-looking individuals now appeared *less* attractive, although the effect on attractiveness was smaller than the effect on trustworthiness judgments. Importantly, as shown in Figure 2, when faces were manipulated to appear trustworthy by the subtraction model, they were also rated as approachable ($B = 0.28$, $SE = 0.01$, 95% CI [0.27, 0.29], $t = 41.17$, $p < .001$) and appearing happy (vs. angry) ($B = 0.35$, $SE < 0.01$, 95% CI [0.34, 0.36], $t = 69.56$, $p < .001$).

The magnitude of the effects for the different ratings suggests that the subtraction-model faces varied more strongly on approachability and emotion ratings than on trustworthiness ratings. Consistent with the literature and Experiment 1, this highlights that attractiveness is indeed important for perceived trustworthiness. Since the manipulated faces were intentionally generated to remove the natural covariance between attractiveness and perceived trustworthiness, our ability to vary trustworthiness independently of attractiveness is limited within this set of faces.

Congruent with the human judgments, estimates of all facial happiness gestures were higher in those faces manipulated to appear “trustworthy” by the subtraction model (AU6 estimate: $B = 0.02$, $SE < 0.01$, 95% CI [0.02, 0.03], $t = 13.75$, $p < .001$; AU12 estimate: $B = 0.04$, $SE < 0.01$, 95% CI [0.04, 0.05], $t = 19.75$, $p < .001$, Happy-expression estimate: $B = 0.002$, $SE < 0.001$, 95% CI [0.001, 0.002], $t = 5.02$, $p < .001$; Figure 3 middle).

With regards to the post hoc analysis of Experiment 1, the same pattern of results was found with the face images (i.e., face images manipulated by the standard trustworthiness model). Across all estimates, faces made to appear trustworthy indeed had facial information related to stronger happy expressions (AU6 estimate: $B = 0.01$, $SE < 0.01$, 95% CI [0.01, 0.02], $t = 7.25$, $p < .001$; AU12 estimate: $B = 0.03$, $SE < 0.01$, 95% CI [0.02, 0.03], $t = 13.95$, $p < .001$; Happy-expression estimate: $B = 0.003$, $SE < 0.001$, 95% CI [0.002, 0.004], $t = 7.81$, $p < .001$; Figure 3 top). This finding shows that smile-related information varied in the original model, consistent with behavioral studies (Oosterhof & Todorov, 2008; Todorov, 2008). However, this result may be still influenced by attractiveness (a face may appear happier/less happy because they are more attractive/unattractive), which is why we needed to test the modified model of trustworthiness (e.g., the subtraction model in Experiment 2).

Overall, these results show that the subtraction model of trustworthiness judgments could make faces appear trustworthy or untrustworthy without relying on attractiveness. It made faces appear trustworthy or untrustworthy through a route separate from the halo effect, namely, facial cues for approachability and happiness, shown via human judgments as well as machine-extracted features.

Experiment 3

Experiment 2 showcased faces that could be manipulated on perceived trustworthiness in the absence of the halo effect. This was done by building a novel face model of trustworthiness judgments (subtraction model) that removed attractiveness cues by subtracting a model of attractiveness. However, the new model manipulated the faces on attractiveness in the opposite direction to trustworthiness judgments (Figure 2). That is, when faces were made to appear more trustworthy, they now appeared *less* attractive, albeit to a smaller degree than the effect on trustworthiness judgments.

This emergent negative correlation between trustworthiness judgments and attractiveness judgments may be undesirable for several reasons. First, the negative correlation may have an unexpected effect on judgments of approachability and emotional expressions: it may be exaggerating the impact of facial cues unrelated to attractiveness on trustworthiness judgments. Second, a researcher interested in identifying the causal effects of perceived trustworthiness in the absence of attractiveness may want to use images that are uncorrelated rather than negatively correlated with attractiveness. This is because a negative correlation may suggest an alternative explanation to the observed results, though one in the opposite direction of the halo effect.

To minimize this negative correlation, Experiment 3 introduces a new model of trustworthiness judgments that is *not* correlated with (rather than *negatively* correlated with, as in the subtraction model) the model of attractiveness ('orthogonal model'). As in Experiment 2, we first tested whether the new model of trustworthiness judgments could manipulate faces to appear more or less trustworthy without affecting attractiveness judgments in the same direction. We also tested, as in Experiment 2, whether the faces manipulated to appear more trustworthy are also perceived as more approachable and happier. We again conducted additional analyses using ML algorithms to investigate the changes in stimulus features related to emotional gestures. Specifically, we predicted smile-related machine-extracted estimates of facial features from the model manipulation level. A positive relationship would show that perceived trustworthiness is related to "smily" facial cues.

Method

Participants. Participants participated in an online study for monetary compensation via Amazon Mechanical Turk. In each condition, we used sample sizes similar to the ones in Experiment 2. Similar to the effect of the subtraction model in Experiment 2, the effect of the "orthogonal" model of trustworthiness judgments was expected to be smaller than that of the standard model (Oh et al., 2019). Four separate groups of participants judged individuals depicted in face images on four different social dimensions: trustworthiness ($n = 57$; M age = 40.02, SD age = 12.47; 23 female, 34 male; 6 Asian, 6 Black, 45 White), attractiveness ($n = 59$; M age = 39.53; SD age = 11.32; 30 female, 29 male; 1 Native American, 3 Asian, 7 Black, 48 White), approachability ($n = 49$, M age = 41.57, SD age = 14.03; 18 female, 31 male; 6 Asian, 6 Black, 36 White, 1 other), and emotionality ($n = 46$, M age = 36.67, SD age = 9.82; 13 female, 32 male, 1 non-binary; 3 Asian, 13 Black, 29 White, 1 other).

Stimuli. We built yet another model of trustworthiness judgments and generated a new set of face images. Instead of subtracting the attractiveness model from the trustworthiness-judgment model as in Experiment 2, we regressed the attractiveness-related information out of the trustworthiness-judgment model. Each model here consists of 100 parameters. One can run a linear regression predicting the trustworthiness-judgment model from the attractiveness model and retain the residuals as the new model parameters. The resulting model represents facial information associated with trustworthiness judgments but orthogonal to facial information associated with attractiveness judgments. To generate facial images, we used the faces of 25 different identities as we did in the previous experiments. Each identity was projected at -3 , -2 , -1 , 0 , 1 , 2 , and 3 SD on the dimension of the orthogonal model ([trustworthiness \perp attractiveness]). The final stimulus set consisted of 175 face images (25 identities * 7 manipulation levels), as in the previous experiments. See Figure 4 for a sample identity varying on the orthogonal dimension.

Procedure. The study followed the same procedures as in Experiment 2. We excluded unreliable participants with within-rater reliability that is less than or equal to 0 from further analyses: 10, 4, 4, and 3 participants from the trustworthiness, attractiveness, approachability, and emotion ratings conditions, respectively. We had 47, 55, 45, and 43 final participants in

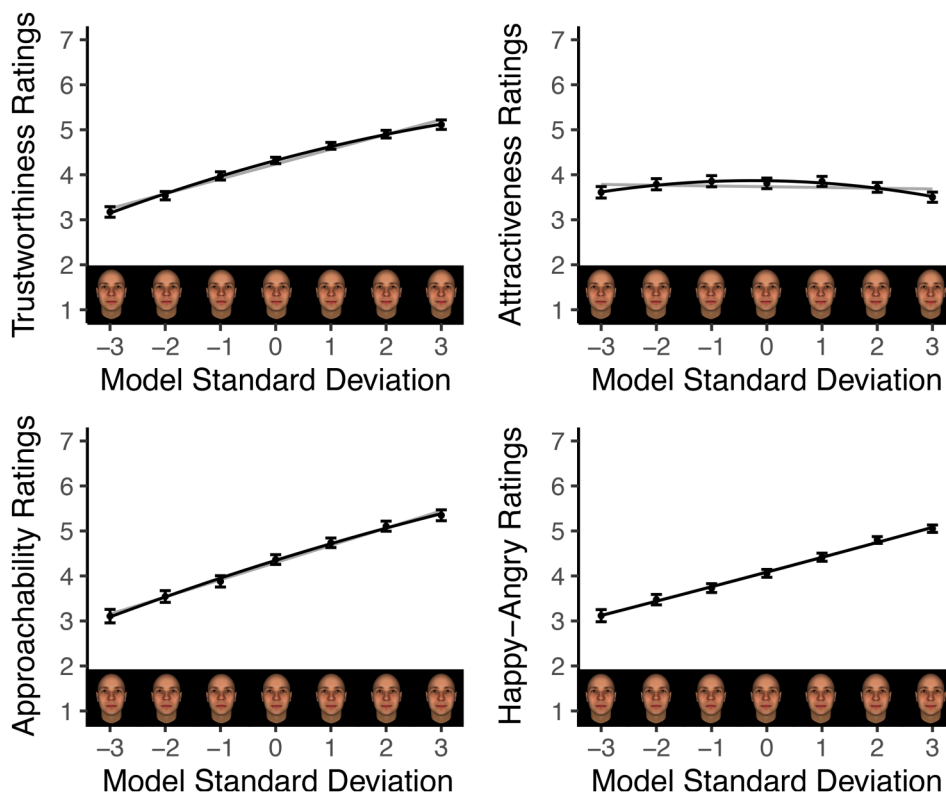


Figure 4. Judgments of trustworthiness, attractiveness, approachability, and emotional expressions of faces manipulated by a statistical model of perceived trustworthiness orthogonal to a model of attractiveness (Experiment 3). To remove the halo effect of attractiveness, we built a new model of trustworthiness judgments that is orthogonal to the model of attractiveness (see main text for details). The orthogonal model of trustworthiness judgments could make faces appear more or less trustworthy (top left). Notably, like the subtraction model (Figure 2), trustworthy-looking faces generated by the orthogonal model were not more attractive (top right). These trustworthy-looking faces were rated to appear approachable (bottom left) and happy (bottom right). These findings suggest that facial information related to approachability and emotional expressions was preserved in the faces, even in the absence of attractiveness cues. The lines denote the linear and quadratic fit across all data points. The error bars denote the standard errors across participants. The lines and error bars are for visualization only; the actual analyses were conducted using mixed-effects models to consider the idiosyncrasies of the participants and face identities. Sample faces originating from one identity are displayed at the bottom of each subplot above the model manipulation level.

each group. Including all participants' data did not change the results (see Supplemental Materials for details). Across raters, we found a high level of consensus in all types of judgments: Trustworthiness ($ICC = 0.95$, Cronbach's $\alpha = 0.95$), attractiveness ($ICC = 0.89$, $\alpha = 0.93$), approachability ($ICC = 0.95$, $\alpha = 0.96$), and emotion ratings ($ICC = 0.95$, $\alpha = 0.97$). As in previous experiments, the level of consensus was high even when all participants were included (Supplemental Tables 1 and 2).

Analysis. We ran mixed-effects regressions with cross-random factors. The models had the same structure as those in the previous experiments. As in Experiment 2, we conducted additional analyses entailing three algorithm-derived emotion estimates from face images: estimates of AU6, AU12, and happy gesture. We predicted these emotion-related estimates from the model manipulation level via multilevel modeling.

Results and Discussion

The orthogonal [trustworthiness \perp attractiveness] model could manipulate faces to appear more or less trustworthy in the expected direction ($B = 0.33$, $SE = 0.01$, 95% $CI [0.32, 0.34]$, $t = 51.83$, $p < .001$; Figure 4). Importantly, attractiveness did not increase as faces were manipulated to appear more trustworthy ($B = -0.02$, $SE = 0.01$, 95% $CI [-0.03, -0.01]$, $t = -2.85$, $p = .004$). Rather, as in Experiment 2, trustworthy-looking faces now appeared *less* attractive, although the effect on attractiveness was much smaller than (a) the effect on trustworthiness judgments and, more importantly, (b) the effect on attractiveness judgments in Experiment 2.

When the faces were manipulated to appear more trustworthy by the orthogonal model, they were also rated as more approachable ($B = 0.38$, $SE = 0.01$, 95% $CI [0.37, 0.40]$, $t = 57.56$, $p < .001$) and as expressing happiness (as opposed to anger) ($B = 0.33$, $SE = 0.01$, 95% $CI [0.32, 0.34]$, $t = 61.07$, $p < .001$).

Congruent with the human judgments, estimates of the three facial happiness gestures were higher in those faces made to appear “trustworthy” using the new, orthogonal model (AU6 estimate: $B = 0.02$, $SE < 0.01$, 95% $CI [0.01, 0.02]$, $t = 11.33$, $p < .001$; AU12 estimate: $B = 0.04$, $SE < 0.01$, 95% $CI [0.03, 0.04]$, $t = 16.62$, $p < .001$, Happy-expression estimate: $B = 0.002$, $SE < 0.01$, 95% $CI [0.001, 0.002]$, $t = 8.92$, $p < .001$) (Figure 3 bottom).

These results show that the orthogonal model of trustworthiness judgments could make faces appear trustworthy or untrustworthy without relying on attractiveness. Replicating Experiment 2, these findings also suggest that the model made faces appear trustworthy or untrustworthy through a route separate from the halo effect, namely, facial “cues” for approachability and happiness, verified by human judgments and machine-extracted estimates.

General Discussion

Intuitive judgments of trustworthiness of others have serious consequences (Todorov et al., 2015). To understand the processes underlying these judgments and their consequences, it is important to identify the visual facial cues used in the judgments. A subset of these visual cues is related to attractiveness (Dion et al., 1972; Eagly et al., 1991), as individuals with attractive faces are often judged as more trustworthy (e.g., Langlois et al., 2000; Todorov, 2008; Todorov et al., 2013). This leaves open the question of whether there are other equally or more important cues used in judgments of trustworthiness. This is particularly important, given that “trustworthiness judgments” are considered a good proxy for the valence evaluation of faces (Oosterhof & Todorov, 2008; Todorov, 2008). Valence consistently emerges as the core dimension of face evaluation across different data-summarizing techniques, samples of face images, and participants’ samples (Jones et al., 2021; Lin et al., 2021; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov & Oh, 2021). Valence evaluation is strongly intertwined with judgments of various traits, including approachability, warmth, and trustworthiness (Jones & Kramer, 2021; Oosterhof & Todorov, 2008; Sutherland et al., 2013; Todorov & Oh, 2021). For example, in the original work positing valence as the primary dimension of face evaluation (Oosterhof & Todorov, 2008), trustworthiness judgments were highly correlated ($>.90$) with valence, which was estimated as a linear combination of multiple social judgments (see Figure S3 in Oosterhof & Todorov, 2008).

Thus, judgments of trustworthiness are a good starting point to study the cues driving general valence evaluation of faces. At the same time, attractiveness is often thought to be the most important attribute of face evaluation, so it is theoretically important to ask to what extent overall valence (not just trustworthiness judgments) from faces can be changed without relying on the halo effect of attractiveness. Practically, being able to manipulate the perceived valence of faces while controlling for the attractiveness halo would make it possible to draw clear causal inferences about the impact of face valence on human decisions.

To answer these questions, using data-driven face models of perceived trustworthiness (Todorov & Oh, 2021), we removed the covarying attractiveness information from the variance in trustworthiness judgments. We accomplished this by building two different models: the first forcing attractiveness and trustworthiness judgments to be negatively correlated by subtracting a model of attractiveness from a model of perceived trustworthiness (Experiment 2) and the second forcing the model of perceived trustworthiness to be orthogonal to the model of perceived attractiveness (Experiment 3). In both cases, faces manipulated to appear more trustworthy were indeed perceived to be more trustworthy but *less* attractive. Notice that any effects of perceived trustworthiness (or general valence) on experimental outcomes using these stimuli cannot be attributed to the attractiveness halo.

More importantly, we identified cues that systematically contribute to trustworthiness judgments (and potentially general valence evaluation) and are independent of attractiveness cues: how approachable and happy a face looks. Specifically, in the absence of the attractiveness halo, perceptions of approachability and happy expressions drove trustworthiness judgments. In other words, we found that faces manipulated to appear more trustworthy were indeed perceived as more approachable and happier. The latter finding was confirmed by both human judgments and ML estimates of facial gestures, indicating happy expressions.

Todorov (2008) has argued that in the absence of clear emotional signals, trustworthiness judgments are an attempt to infer whether to approach or avoid a person based on the similarity of the person's facial features to emotional expressions signaling approach or avoidance behaviors. The current findings are consistent with this argument and add to prior research that highlights the importance of smiling (implied via face structure, e.g., upturned lips) in valence evaluation as well as in many specific valence-related impressions (Jaeger & Jones, 2021; Jones & Kramer, 2021; Lin et al., 2021; Oosterhof & Todorov, 2008; Peterson et al., 2022; Sutherland et al., 2013). Indeed, consistent with the notion that valence evaluation from a resting, "neutral" face is ultimately an attempt to decide whether to approach or avoid an individual (Todorov, 2008), a stable "trustworthy facial look" and an emotional expression engage the same perceptual mechanism in humans (Engell et al., 2010; Oosterhof & Todorov, 2009; Said et al., 2009).

However, the current findings reveal that multiple facial cues are at the basis of trustworthiness (or general valence) judgments. At a minimum, these include cues for attractiveness, emotions of happiness, and approachability. Note that our computational approach can be used to further isolate specific cues. Undoubtedly, as indicated by the findings (Figures 2–4), emotional cues and approachability cues are redundant. However, it is straightforward to build models of perceived approachability and positive emotions, control for the redundancy of cues, and test hypotheses for even more specific sets of cues. For example, facial cues associated with femininity and masculinity are likely associated with perceived approachability irrespective of emotional cues.

The current findings also highlight the intercorrelated nature of trait judgments. Judgments of social traits (e.g., trustworthiness, competence) and other attributes (e.g., age, masculinity) from faces are correlated with each other to various degrees (Berry & Zebrowitz-McArthur, 1986; Johnson et al., 2012; Jones et al., 2021; Todorov & Oh, 2021), and the halo effect of facial attractiveness is just one—albeit prominent—example. The pattern of judgment associations stems from concepts people have about the judged attributes (e.g., a belief about how an individual with a trait, such as friendliness, also has another trait, such as adventurousness) (Stolier et al., 2018), and can be acquired through learning (Oh et al., 2022; Stolier et al., 2020). However, despite the intercorrelated nature of social judgments from faces, the present approach shows that it is possible to systematically dissect the complex associative web of face judgments. Multiple visual facial "cues" are at the basis of any single social judgment. But within a modeling framework (e.g., Blanz & Vetter, 1999; Oosterhof & Todorov, 2008; Peterson et al., 2022; Walker & Vetter, 2009), one can isolate and then quantitatively manipulate specific sets of facial cues associated with the judgment.

As part of our studies, we generated three sets of new face images (525 images in total = 25 face identities * 7 manipulation levels * 3 trustworthiness models). They are images of 25 identities manipulated by (1) the “standard” trustworthiness model (in which faces varied both on perceived trustworthiness and attractiveness in the same direction), (2) the “subtraction” model (in which faces varied on perceived trustworthiness in the opposite direction to attractiveness), and (3) the “orthogonal” model (in which faces varied on perceived trustworthiness judgments but varied little on perceived attractiveness). We hope that these stimulus sets are useful for researchers interested in the effect of perceived trust on human behavior. The sets include a collection of faces that vary in face-based judgments of trustworthiness but are free of (or negatively correlated with) the attractiveness halo effect. These stimuli should make it possible to test for causal effects of superficial trustworthiness judgments from human faces on social interactions and outcomes *in the absence of* the halo effect.

One limitation of the current studies is that we used synthetic face images, whose processing in the human mind may be different from the processing of real-life face images (Balas & Pacella, 2015, 2017). Due to ambiguity in gender, for example, the face images used here do not allow to explore the role of facial gender differences, a key factor in social perception (Mileva et al., 2019; Oh, Dotsch et al., 2020; Sutherland et al., 2015). Specifically, “face trustworthiness” and attractiveness may have a nonlinear relationship for female faces (Sofer et al., 2015), and are more strongly correlated with each other for male than for female faces (Mileva et al., 2019). However, the computational approach described here is easily extendable to both hyper-realistic synthetic and real-life facial images (Peterson et al., 2022), overcoming these shortcomings of the present studies.

One may question the practical implications of the data reported here to the extent that it may be impossible to change one’s face “trustworthiness” independently of attractiveness in real life. However, we note two things. First, the primary objective of this work was to identify the specific cues underlying judgments of trustworthiness (and possibly valence evaluation) while controlling for the attractiveness halo. As we noted above, the findings have both theoretical and practical implications (e.g., providing stimuli for experiments that can draw clear causal inferences about the influence of specific judgments on behaviors). Second, the findings clearly show that emotional expressive behavior is an important cue for perceived trustworthiness. This is consistent with work suggesting that facial emotions can overwrite the influence of morphological features on impressions (Gill et al., 2014).

In conclusion, the present study shows that it is possible to isolate visual cues associated with attractiveness from visual cues associated with perceived trustworthiness (and potentially general valence evaluation of faces). The latter cues are associated with emotional expressions and perceived approachability. The approach outlined here can be used to further dissect these cues. This work adds to the growing literature on the multifaceted basis of social facial perception.

Acknowledgements

The authors thank Darrel J. H. Chan and Justina S. C. Tan for helping with analysis.

Author Contribution(s)

DongWon Oh: Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing – original draft; Writing – review & editing.

Nicole Wedel: Conceptualization; Investigation; Methodology.

Brandon Labbree: Investigation.

Alexander Todorov: Conceptualization; Methodology; Resources; Supervision; Writing – review & editing.

Declaration of Conflicting Interests


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business.

ORCID iDs

DongWon Oh  <https://orcid.org/0000-0002-2105-3756>

Brandon Labbree  <https://orcid.org/0000-0002-5893-4469>

Supplemental Material

Supplemental material for this article is available online.

References

- Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, *393*, 470–474. <https://doi.org/10.1038/30982>
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Balas, B., & Pacella, J. (2015). Artificial faces are harder to remember. *Computers in Human Behavior*, *52*, 331–337. <https://doi.org/10.1016/j.chb.2015.06.018>
- Balas, B., & Pacella, J. (2017). Trustworthiness perception is disrupted in artificial faces. *Computers in Human Behavior*, *77*, 240–248. <https://doi.org/10.1016/j.chb.2017.08.045>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Berry, D. S., & Zebrowitz-McArthur, L. (1986). Perceiving character in faces: The impact of age-related craniofacial changes on social perception. *Psychological Bulletin*, *100*, 3–18. <https://doi.org/10.1037/0033-2909.100.1.3>
- Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, *19*, 187–194. <https://doi.org/10.1145/311535.311556>
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system [Paper presented]. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.
- Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*, 285–290. <https://doi.org/10.1037/h0033731>
- Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies*, *25*, 2455–2484. <https://doi.org/10.1093/rfs/hhs071>
- Eagly, A. H., Ashmore, R. D., & Makhijani, M. G. (1991). What is beautiful is good, but...: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*, 109–128. <https://doi.org/10.1037/0033-2909.110.1.109>
- Engell, A. D., Todorov, A., & Haxby, J. V. (2010). Common neural mechanisms for the evaluation of facial trustworthiness and emotional expressions as revealed by behavioral adaptation. *Perception*, *39*, 931–941. <https://doi.org/10.1068/p6633>
- Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in Airbnb. *Tourism Management*, *55*, 62–73. <https://doi.org/10.1016/j.tourman.2016.01.013>
- Friesen, E., & Ekman, P. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists.

- Gill, D., Garrod, O. G. B., Jack, R. E., & Schyns, P. G. (2014). Facial movements strategically camouflage involuntary social signals of face morphology. *Psychological Science*, *25*, 1079–1086. <https://doi.org/10.1177/0956797614522274>
- Gomulya, D., Wong, E. M., Ormiston, M. E., & Boeker, W. (2017). The role of facial appearance on CEO selection after firm misconduct. *Journal of Applied Psychology*, *102*, 617–635. <https://doi.org/10.1037/apl0000172>
- Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Jaeger, B., & Jones, A. L. (2021). Which facial features are central in impression formation? *Social Psychological and Personality Science*, *13*, 553–561. <https://doi.org/10.1177/19485506211034979>
- Johnson, K. L., Freeman, J. B., & Pauker, K. (2012). Race is gendered: How covarying phenotypes and stereotypes bias sex categorization. *Journal of Personality and Social Psychology*, *102*, 116–131. <https://doi.org/10.1037/a0025335>
- Jolly, E., Cheong, J. H., Xie, T., Byrne, S., Kenny, M., & Chang, L. J. (2021). *Py-feat: Python facial expression analysis toolbox*. arXiv. arXiv.
- Jones, A. L., & Kramer, R. S. S. (2021). Facial first impressions form two clusters representing approach-avoidance. *Cognitive Psychology*, *126*, 101387. <https://doi.org/10.1016/j.cogpsych.2021.101387>
- Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxson, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., ... N. A. Coles (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159–169. <https://doi.org/10.1038/s41562-020-01007-2>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. <https://doi.org/10.18637/jss.v082.i13>
- 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*, 390–423. <https://doi.org/10.1037/0033-2909.126.3.390>
- Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterizing trait attributions from faces. *Nature Communications*, *12*, 5168. <https://doi.org/10.1038/s41467-021-25500-y>
- Mileva, M., Kramer, R. S. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception*, *48*, 471–486. <https://doi.org/10.1177/0301006619848996>
- Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, *30*, 65–79. <https://doi.org/10.1177/0956797618813092>
- Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General*, *149*, 323–342. <https://doi.org/10.1037/xge0000638>
- Oh, D., Grant-Villegas, N., & Todorov, A. (2020). The eye wants what the heart wants: Females' preference in male faces are related to partner personality preference. *Journal of Experimental Psychology: Human Perception and Performance*, *46*, 1328–1343. <https://doi.org/10.1037/xhp0000858>
- Oh, D., Martin, J. D., & Freeman, J. B. (2022). Personality across world regions predicts variability in the structure of face impressions. *Psychological Science*, *33*, 1240–1256. <https://doi.org/10.1177/09567976211072814>
- Olivola, C. Y., & Todorov, A. (2017). The biasing effects of appearances go beyond physical attractiveness and mating motives. *Behavioral and Brain Sciences*, *40*, e38. <https://doi.org/10.1017/S0140525X16000595>
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*, 11087–11092. <https://doi.org/10.1073/pnas.0805664105>
- Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*, 128–133. <https://doi.org/10.1037/a0014520>
- Peterson, J. C., Uddenberg, S., Griffiths, T. L., Todorov, A., & Suchow, J. W. (2022). Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, *119*, e2115228119. <https://doi.org/10.1073/pnas.2115228119>
- Pham, L., Vu, T. H., & Tran, T. A. (2021). Facial expression recognition using residual masking network [Paper presented]. 25th International Conference on Pattern Recognition.

- Porter, S., ten Brinke, L., & Gustaw, C. (2010). Dangerous decisions: The impact of first impressions of trustworthiness on the evaluation of legal evidence and defendant culpability. *Psychology, Crime & Law, 16*, 477–491. <https://doi.org/10.1080/10683160902926141>
- R Core Team. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org>
- Said, C. P., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion, 9*, 260–264. <https://doi.org/10.1037/a0014681>
- Singular Inversions. FaceGen software development kit. Toronto, Canada.
- Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychological Science, 26*, 39–47. <https://doi.org/10.1177/0956797614554955>
- Stewart, J. E. II. (1980). Defendant's attractiveness as a factor in the outcome of criminal trials: An observational study. *Journal of Applied Social Psychology, 10*, 348–361. <https://doi.org/10.1111/j.1559-1816.1980.tb00715.x>
- Stewart, J. E. II. (1985). Appearance and punishment: The attraction-lenience effect in the courtroom. *The Journal of Social Psychology, 125*, 373–378. <https://doi.org/10.1080/00224545.1985.9922900>
- Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour, 4*, 361–371. <https://doi.org/10.1038/s41562-019-0800-6>
- Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences, 115*, 9210–9215. <https://doi.org/10.1073/pnas.1807222115>
- Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*, 105–118. <https://doi.org/10.1016/j.cognition.2012.12.001>
- Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology, 106*, 186–208. <https://doi.org/10.1111/bjop.12085>
- Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25–29. <https://doi.org/10.1037/h0071663>
- Todorov, A. (2008). Evaluating faces on trustworthiness: An extension of systems for recognition of emotions signaling approach/avoidance behaviors. *Annals of the New York Academy of Sciences, 1124*, 208–224. <https://doi.org/10.1196/annals.1440.012>
- Todorov, A., Dotsch, R., Porter, J. M., Oosterhof, N. N., & Falvello, V. B. (2013). Validation of data-driven computational models of social perception of faces. *Emotion, 13*, 724–738. <https://doi.org/10.1037/a0032335>
- Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. 63, pp. 189–245). Academic Press.
- 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. <https://doi.org/10.1146/annurev-psych-113011-143831>
- Valentine, T. (2001). Face-space models of face recognition. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 83–113). Lawrence Erlbaum Associates.
- Walker, M., & Vetter, T. (2009). Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision, 9*, 1–13. <https://doi.org/10.1167/9.11.12>
- Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science, 26*, 1325–1331. <https://doi.org/10.1177/0956797615590992>