DOI: 10.1111/nyas.15293

CONCISE REVIEW

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Face evaluation: Findings, methods, and challenges

Alexander Todorov[1] | DongWon Oh[2] | Stefan Uddenberg[3] | Daniel N. Albohn[1]

[1]The University of Chicago Booth School of Business, Chicago, Illinois, USA

[2]Department of Psychology, National University of Singapore, Singapore, Singapore

[3]Department of Psychology, University of Illinois Urbana-Champaign, Champaign, Illinois, USA

Correspondence
Alexander Todorov, The University of Chicago Booth School of Business, Chicago, IL, USA.
Email: alextodorov@uchicago.edu

Funding information
Booth School of Business, University of Chicago

Abstract

Complex evaluative judgments from facial appearance are made efficiently and are consequential. We review some of the most important findings and methods over the last two decades of research on face evaluation. Such evaluative judgments emerge early in development and show a surprising consistency over time and across cultures. Judgments of trustworthiness, in particular, are closely associated with general valence evaluation of faces and are grounded in resemblance to emotional expressions, signaling approach versus avoidance behaviors. Data-driven computational models have been critical for the discovery of the configurations of features, including resemblance to emotional expressions, driving specific judgments. However, almost all models are based on judgments aggregated across individuals, essentially masking idiosyncratic differences in judgments. Yet, recent research shows that most of the meaningful variance of complex judgments such as trustworthiness is idiosyncratic: explained not by stimulus features, but by participants and participants by stimuli interactions. Hence, to understand complex judgments, we need to develop methods for building models of judgments of individual participants. We describe one such method, combining the strengths of well-established methods with recent developments in machine learning.

KEYWORDS
data-driven computational methods, face evaluation, judgment

More than 15 years ago, we introduced data-driven computational models for visualizing complex social judgments from faces.[1,2] The objective of these methods was to identify the perceptual features that drive specific judgments or read the mental representations underlying these judgments. Our earlier manuscript "Evaluating faces on trustworthiness" (Todorov, 2008)[1] was focused on substantive findings about the nature of trustworthiness judgments. Perhaps the most important findings were identifying these judgments as a proxy for general valence evaluation of faces (i.e., good versus bad) and the close relationship between this evaluation and emotional expressions, signaling approach versus avoidance behavior. As outlined in the first section ("Complex judgments from faces"), these findings, as well as the findings about the efficiency of trustworthiness judgments (e.g., made rapidly from minimal information with little effort), have withstood the test of time rather well.[3,4] Moreover, these judgments turned out to be remarkably consistent across time and cultures.[5]

The last section of Todorov (2008) was on the advantages of building data-driven computational models of complex judgments. With hindsight, this line of work has been the most generative. Although there were some early attempts to model social perception[6] and certainly many related methods in psychophysics,[7–11] this methodological approach was not firmly established in the domain of complex social judgments. In contrast to standard, theory-driven approaches, this approach allows for the discovery of configurations of features that

drive complex judgments, without imposing any prior assumptions about what features matter or not.[12] The methods were developed by Todorov and Oosterhof[2,13] and have undergone considerable development over time, as outlined in the section below "Data-driven computational methods for modeling social judgments." This section also describes the remarkable recent developments in the field, following the introduction of deep neural nets and generative adversarial networks (GANs).

One development that was not foreseen in Todorov (2008) was the importance of idiosyncratic differences in face evaluation. Although there were singular voices drawing attention to the importance of these differences,[14] idiosyncratic differences were largely overlooked until recently. However, as it turned out, these differences explain most of the meaningful variance of complex judgments such as trustworthiness.[15,16] This finding has dramatic implications for how face evaluation should be modeled. The section "The importance of idiosyncratic differences in face evaluation" outlines recent work on identifying idiosyncratic and shared contributions to judgments from faces and new methods for building idiosyncratic models.

COMPLEX JUDGMENTS FROM FACES

People efficiently extract information from faces to infer not only attributes that can be read from the face such as age and sex,[17] but also attributes that are read into the face such as perceived trustworthiness and competence.[18–23] Typically, in these studies, faces are presented briefly and the criterion is the judgment people make in the absence of time constraints. For attributes that can be read from faces (e.g., age), exposures of 50 ms are sufficient for people to make judgments that almost perfectly approximate their judgments made in the absence of time constraints.[17] For attributes that are read into faces (e.g., perceived trustworthiness), these exposures are in the order of 150−200 ms. Note that although an individual could be highly consistent in their own judgments, indicating high intraindividual consistency, they may be highly inconsistent with judgments of other individuals, indicating low interindividual consistency.[16,24] In fact, as shown in the section "The importance of idiosyncratic differences in face evaluation" below, complex judgments from faces tend to be highly idiosyncratic.[15]

We focus here on judgments of trustworthiness, because this was the focus of the paper in 2008,[1] but the findings and methods generalize to other complex judgments. Besides the findings that these judgments are made after minimal exposure to faces, several other findings are notable. First, although most of the findings described above have been observed when people were asked to explicitly judge faces, explicit intention is not necessary to document the effects of perceived facial trustworthiness.[25–29] Recent studies using fast periodic visual stimulation have been particularly informative for the study of face perception.[30] In this approach, faces are presented at a fixed, periodic rate. This presentation evokes detectable corresponding periodic changes in the voltage amplitude measured on the scalp with electroencephalography (EEG). Contrasting two conditions (e.g., types of faces) at the same rate can identify whether the brain is discriminating between these two conditions. The measured response has a high signal-to-noise ratio relative to standard EEG measures and is objective because the frequency is explicitly defined by the experimenter. In one of the first studies using this technique to study perceived facial trustworthiness, Verosky and colleagues[26] presented faces at a rate of 6 Hz (about 167 ms) and also included oddball faces mismatched on perceived trustworthiness. They found consistent and widespread neural responses to the perceived trustworthiness of the oddball faces, although the participants' task did not involve any evaluation of the faces (their task was to attend to the color of a fixation cross in the middle of the screen and detect changes in this color). Subsequent studies also showed a reliable neural sensitivity to facial trustworthiness in tasks not requiring judgments of trustworthiness[27] and, in fact, this sensitivity was not modulated by task instructions.[28]

The second notable finding is that trustworthiness judgments emerge early in development.[29,31–37] Three- to four-year-old children make trustworthiness judgments, which are similar to adults' judgments,[32] and even 7-month-old infants appear to be sensitive to differences in perceived facial trustworthiness, although not perceived facial dominance.[37]

The third notable finding is that trustworthiness judgments aggregated across individuals are highly consistent over time. We collected judgments of the same faces from different samples of participants more than 10 years apart.[5,38] Nonetheless, as shown in Figure 1A, the judgments were highly correlated ($r = 0.88$). Fourth and perhaps more surprisingly, trustworthiness judgments are highly consistent across cultures. A large study collected judgments of the same faces in 11 different world regions.[39] As shown in Figure 1B, trustworthiness judgments in different regions were highly correlated, with the correlations ranging from 0.71 to 0.96.

The main reason for the early focus on trustworthiness judgments was that they were highly correlated with almost any other judgment with an evaluative component (e.g., good versus bad). In principal component and factor analyses of social judgments from faces, the first component invariably captures valence evaluation of faces,[2,40,41] and this component is highly correlated with judgments of trustworthiness, even when these judgments are not part of the initial input to the analyses.[2,42] As shown in Figure 1C, this high correlation between trustworthiness judgments and valence evaluation, estimated from a linear combination of 12 other social judgments, replicates across world regions. The median correlation is 0.92, with a range from 0.75 to 0.96. These findings support the early arguments that in the absence of a specific context, trustworthiness judgments are a proxy for a general valence evaluation of faces and that this evaluation is in the service of approach versus avoidance decisions.[1,42] In fact, studies that rely on unsupervised clustering of faces, based on their social judgments, show two fundamental clusters of faces that map onto the first valence component and are tightly associated with approach versus avoidance decisions.[43]

The finding that faces are clustered according to their perceived approachability nicely dovetails with the very first findings of the data-driven computational models, described in the section "Data-driven

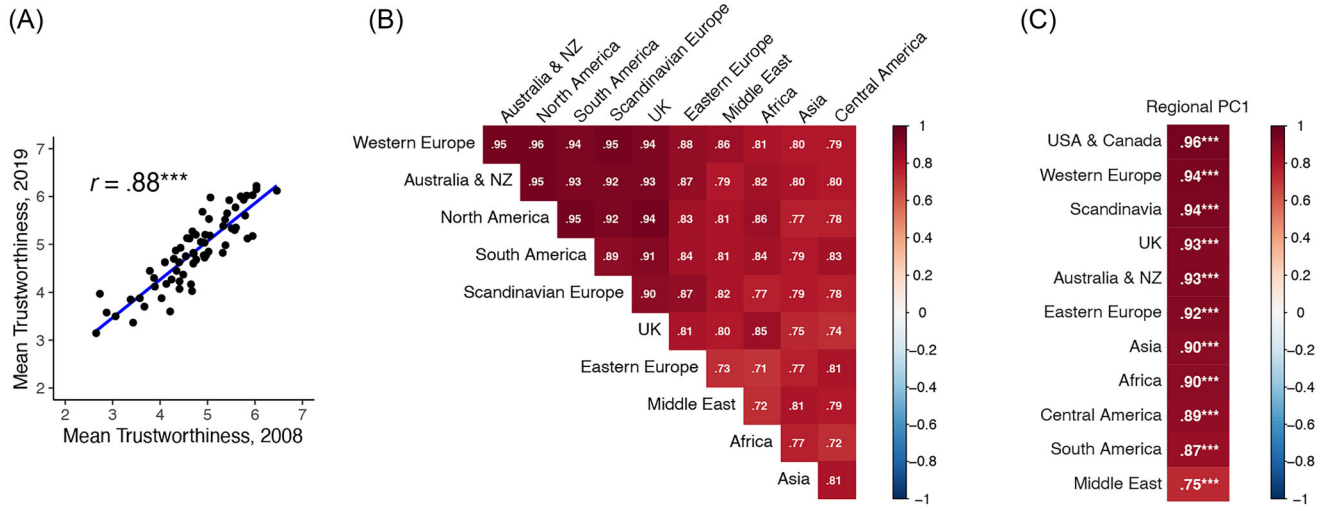

FIGURE 1 The temporal and cross-cultural consistency of trustworthiness judgments from faces. (A) A scatter plot of judgments of faces collected more than 10 years apart from two different samples (data from Oosterhof and Todorov, 2008 and Oh et al., 2019).[2,38] Each point in the scatter plot is a face. (B) All pair-wise correlations of trustworthiness judgments in 11 world regions (data from Jones et al., 2021).[39] (C) Correlations between trustworthiness judgments and the first principal component derived from a PCA of 12 other social judgments in 11 world regions (data from Jones et al., 2021).[39] This component is best interpreted as valence evaluation.

FIGURE 2 Data-driven computational models of judgments of trustworthiness. As faces change from left to right, their perceived trustworthiness increases. (A) A model that visualizes face shape information associated with perceived trustworthiness (adapted from Oosterhof and Todorov, 2008).[2] (B) A model that visualizes face shape and reflectance information associated with perceived trustworthiness (adapted from Todorov and Oosterhof, 2011).[13] (C) A model that visualizes face shape and reflectance information associated with perceived trustworthiness while controlling for attractiveness (adapted from Oh et al., 2023).[58]

computational methods for modeling social judgments". Specifically, using a model of trustworthiness judgments (see Figure 2) to exaggerate the features that lead to judgments of untrustworthiness versus trustworthiness resulted in faces expressing anger versus happiness, respectively.[1,2] This was the case even though the input to the model was judgments of emotionally neutral faces. Thus, subtle traces of emotional expressions, signaling approach versus avoidance behavior, are used to make trustworthiness judgments, perhaps explaining the rather surprising sensitivity of infants to perceived facial trustworthiness.[37]

The link between trustworthiness judgments/valence evaluation and emotional expressions has been confirmed in a variety of paradigms. In dynamic morphing studies, emotions congruent with facial features (e.g., smiling and trustworthy features) are perceived as more intense.[44] In behavioral adaptation studies, adapting to angry (versus happy) expressions increases (versus decreases) the trustworthiness evaluation of emotionally neutral faces.[45] In both behavioral and machine learning studies, the resemblance of neutral faces to emotional expressions predicts complex judgments, including trustworthiness.[46–50] Finally, different versions of reverse correlation approaches, in which combinations of facial features are used to predict social judgments show similar links between the latter and emotional expressions, signaling approach versus avoidance behaviors.[43,51–53]

In sum, complex evaluative judgments from facial appearance are made efficiently, irrespective of intentions to evaluate or not, emerge early in development, and show both temporal and cross-cultural consistency, at least when aggregated across participants. One of the key inputs to these judgments is emotional expressions, signaling approach versus avoidance behaviors. Even when the faces appear to be emotionally neutral, their resemblance to specific emotional expressions shapes the evaluative judgments.

DATA-DRIVEN COMPUTATIONAL METHODS FOR MODELING SOCIAL JUDGMENTS

In a standard, theory-driven approach, one starts with a specific hypothesis (e.g., the shape of eyebrows is related to perceived trustworthiness), manipulates the key variables (e.g., eyebrows shape), and observes the effect on judgments (e.g., trustworthiness). Some of the problems with this approach are that (a) the space of hypotheses is infinitely large (20 binary features result in more than 1 million combinations; and features are not binary); (b) it is not clear a priori what qualifies as a feature (e.g., mouth versus corner of a mouth versus pixel); and (c) features that are important for judgments but not in the mind of the experimenter are never studied.[12,51]

In contrast to theory-driven methods, in a data-driven approach, one starts with a random sampling of stimuli from a well-defined space, has these stimuli judged on a specific dimension, and looks for variations in the features, defined in the space, that predict the judgment. There are four principal stages of this approach. First, one needs a statistical representational space of the stimulus domain (e.g., faces) that allows for random sampling of stimuli. This is essential because these methods are a version of reverse correlation, in which the outcome variable (e.g., judgment) is parametrically modeled as a function of the random variation of the stimuli.[12] As described below, this statistical space could be based on a principal components analysis (PCA) of the shape and texture of faces, as in our earlier work,[1,2] or on deep machine learning from thousands of images, as in our recent work.[54] The first stage is randomly sampling stimuli from the representational space. The second stage is the evaluation of the randomly generated stimuli. At this stage, it is essential to establish that the evaluation is statistically reliable. We note that although the typical evaluation procedure entails the rating of images, many other outcome variables could be modeled—from response times to pupil dilation to neuronal responses—as long as the measures are statistically reliable. The third stage is the building of a model of the evaluation in the statistical representational space of the stimulus domain. The final stage is the validation of this model. This stage entails generating novel stimuli, manipulating these stimuli by the model, and having the stimuli evaluated by a novel group of participants. In a successful validation, the manipulated stimuli should be evaluated as intended by the model.[5,55]

In our early work,[1,2] we randomly sampled faces from a 50-dimensional shape space of faces, derived from 3D laser scans of real faces. Participants judged several hundred of these randomly sampled faces on trustworthiness (and also dominance and threat in Oosterhof and Todorov, 2008),[2] and we used the average judgment to find variation in the shape space that predicts changes in judgments (a detailed treatment of these specific methods and their assumptions is provided elsewhere[5]). Figure 2A shows a model of perceived trustworthiness. As mentioned in the section "Complex judgments from faces", one can see that emotional expressions emerge despite the fact that we only used faces that appeared to be completely emotionally neutral. One can also see that trustworthy-looking faces are more feminine and baby-faced, a finding consistent with many prior studies.[56,57]

The first models of complex judgments that we built were models based on facial shape, but facial reflectance (brightness, texture, and color variation) is just as important for these judgments.[59,60] Figure 2B shows a model of perceived trustworthiness that manipulates both shape and reflectance. The influence of masculinity is particularly salient here, as male faces tend to be darker than female faces.[61,62] In subsequent research, we built and validated models of dozens of judgments based on both shape and reflectance.[13,55,63,64] The faces generated by these models have been used by thousands of researchers from hundreds of universities covering the globe.[5]

Having a model allows you to inspect the configurations of features that drive specific judgments and to parametrically manipulate the impressions of any facial image. Furthermore, the fact that the models are vectors in the same space has three important implications. First, the similarity of the models is immediately apparent. Not surprisingly, similar, correlated judgments (e.g., trustworthiness and emotional stability) result in similar models.[55] Second, it is straightforward to control for shared variance between different models.[58,65] Figure 2C shows a model of perceived trustworthiness controlling for attractiveness.[58] Although the perceived trustworthiness of faces increases, their attractiveness does not. However, the emotional expressions of the faces predictably change from angry to happy and, correspondingly, their perceptions of approachability.

The third implication is that one can build models of measures, including neural responses, different from explicit judgments and immediately relate these models to more interpretable models of judgments.[66,67] For example, using a continuous flash suppression procedure, we built a model of the speed of emergence of faces in consciousness.[66] This model was highly correlated with a model of dominance judgments: more dominant-looking faces emerged faster

FIGURE 3 Data-driven computational models of judgments of trustworthiness. As faces change from left to right, their perceived trustworthiness increases (adapted from Peterson et al., 2022).[54] (A) A model applied to a female face. (B) A model applied to a male face.

in consciousness. We want to emphasize that the approach need not be applied to explicit judgments only. As noted earlier, any outcome measure of theoretical interest (e.g., response times, approach behavior, pupil dilation as a measure of arousal, etc.) that is statistically reliable could be modeled. At the same time, the existing and interpretable models of explicit judgments provide meaningful constraints on the interpretation of measures with less clear behavioral meaning.

One issue with the faces generated by our older models is that they are highly unrealistic (see Figure 2), although it is possible to apply them to images of real faces to manipulate the impressions of the latter through morphing.[65,68] However, with the remarkable recent rapid developments in the generation of hyper-realistic images such as in the Style-GAN architecture,[69,70] it is possible to build models of hyper-realistic faces.[54] Although the underlying latent representation of hyper-realistic faces (i.e., the statistical representational space) is much more complicated and more difficult to interpret than the PCA-derived representations derived from laser scans of real faces,[5,71] the conceptual logic of building models of judgments is the same. One starts with a random sample of facial images, these images are judged on specific dimensions, the average judgment is used to build a model of the judgment in the latent multidimensional space representing faces, and the model is validated.

Recently, we built more than 30 models of perceived attributes: from attributes that are read from faces (e.g., age, hair color) to attributes that are read into faces (e.g., perceived trustworthiness).[54] Figure 3A shows a model of perceived trustworthiness applied to a female face and Figure 3B shows the same model applied to a male face. The faces are highly realistic and once again emotional expressions emerge as in the old models of synthetic faces. As the faces are manipulated to appear more trustworthy, their emotional expressions become more positive.

One can also control for shared variance with other judgments. We illustrate this with models of two correlated judgments: electability and dominance. As shown in Figure 4, as faces are manipulated to appear more electable, their perceived dominance also increases. Controlling for the latter, the more electable faces acquire more positive expressions.

In sum, there has been remarkable progress in the development of the data-driven computational approach for modeling complex social judgments from faces. This approach discovers the configurations of perceptual features that drive specific judgments without imposing a priori theoretical assumptions about the importance of any feature. Further, this approach is not limited to explicit judgments and can be extended to any behavioral, physiological, or neural measure, as long as this measure is reliably measured.

THE IMPORTANCE OF IDIOSYNCRATIC DIFFERENCES IN FACE EVALUATION

The models of various judgments have been extensively validated,[5,54] but they are models of aggregated judgments. In general, to the extent that there is any agreement in judgments, aggregation would increase the reliability of the judgments. However, it would also mask stable individual differences. The typical statistic of agreement reported in studies is the Cronbach's alpha, with values often higher than 0.90, but this statistic is best interpreted as the expected correlation between the aggregated judgments of two different samples with the same size. Thus, although this statistic indicates the high reliability of aggregated judgments, it does not imply anything about individual differences in judgments. To identify whether these differences meaningfully contribute to judgments, one needs to use repeated judgments of the same stimuli (e.g., faces) and partition the meaningful variance.[14,16]

In variance partitioning studies, the meaningful variance is attributed to the stimuli (i.e., shared contributions to judgments), the participants, and the participants by stimuli interactions (i.e., idiosyncratic contributions to judgments). How the variance partitions is critical for understanding complex judgments (a detailed treatment of the methods, including scripts for analyses and data simulations with recommendations for sample sizes of both participants and stimuli is provided elsewhere[16]). Consider two possibilities: most

FIGURE 4 Data-driven computational models of judgments of electability. As faces change from left to right, their perceived electability increases (adapted from Peterson et al., 2022).[54] (A) A model applied to a female face. (B) A model applied to the same female face while controlling for perceived dominance. (C) A model applied to a male face. (D) A model applied to the same male face while controlling for perceived dominance.

of the variance in judgments is due to the stimuli versus most of the variance is due to the participants by stimuli interaction (e.g., participant 1 likes face A more than face B, but participant 2 likes face B more than face A). In the former case, relying on a model of aggregated judgments is a prudent approach. But in the latter case, this approach is essentially masking most of the meaningful variance and, as a result, providing a misleading picture of the judgment at hand.

In the case of trustworthiness judgments, as shown in Figure 5A,B, the idiosyncratic variance trumps the shared variance.[15,72] In fact, stimulus features account for less than 10% of the meaningful variance of judgments. This result—idiosyncratic exceeding shared variance—holds for other complex judgments from faces.[15,73–75] The only judgments for which the shared variance trumps idiosyncratic variance are relatively simple judgments such as femininity/masculinity and age.[15] In the case of these judgments, in contrast to complex judgments such as trustworthiness, the mapping from facial features to judgments is relatively consistent across participants.

These findings have dramatic implications for how we should build models of complex judgments. The existing models (Figures 2–4) are essentially models of stimulus features that are consistently used by most participants. But these features account for a small proportion of the variance of judgments. Hence, the models effectively hide the highly heterogeneous nature of judgments. Recently, we introduced a novel method for building models of judgments of individual participants.[72] The method combines procedures from classic psychophysical reverse correlation studies[76] and sampling of faces from a latent multidimensional space.[54] As shown in Figure 5C, the resulting models are compelling and highly diverse.

As in the case of models of aggregated judgments, these individual models need to be validated. We have shown that for complex judgments such as trustworthiness, models derived from judgments of the participants are more predictive of their judgments of novel faces than models derived from judgments of other participants.[78] For simple judgments such as masculinity, the predictive power is the same, justifying the reliance on models of aggregated judgments.

The findings of the highly idiosyncratic nature of complex judgments from faces are consistent with twin studies, showing that these judgments are primarily explained by the unique environmental history of the individual.[79,80] This poses particular difficulties for identifying the source of idiosyncratic differences. In fact, modeling those differences is exceedingly difficult.[15] We can make informed empirical guesses about their source—for example, the cultural typicality of faces and their resemblance to personally familiar

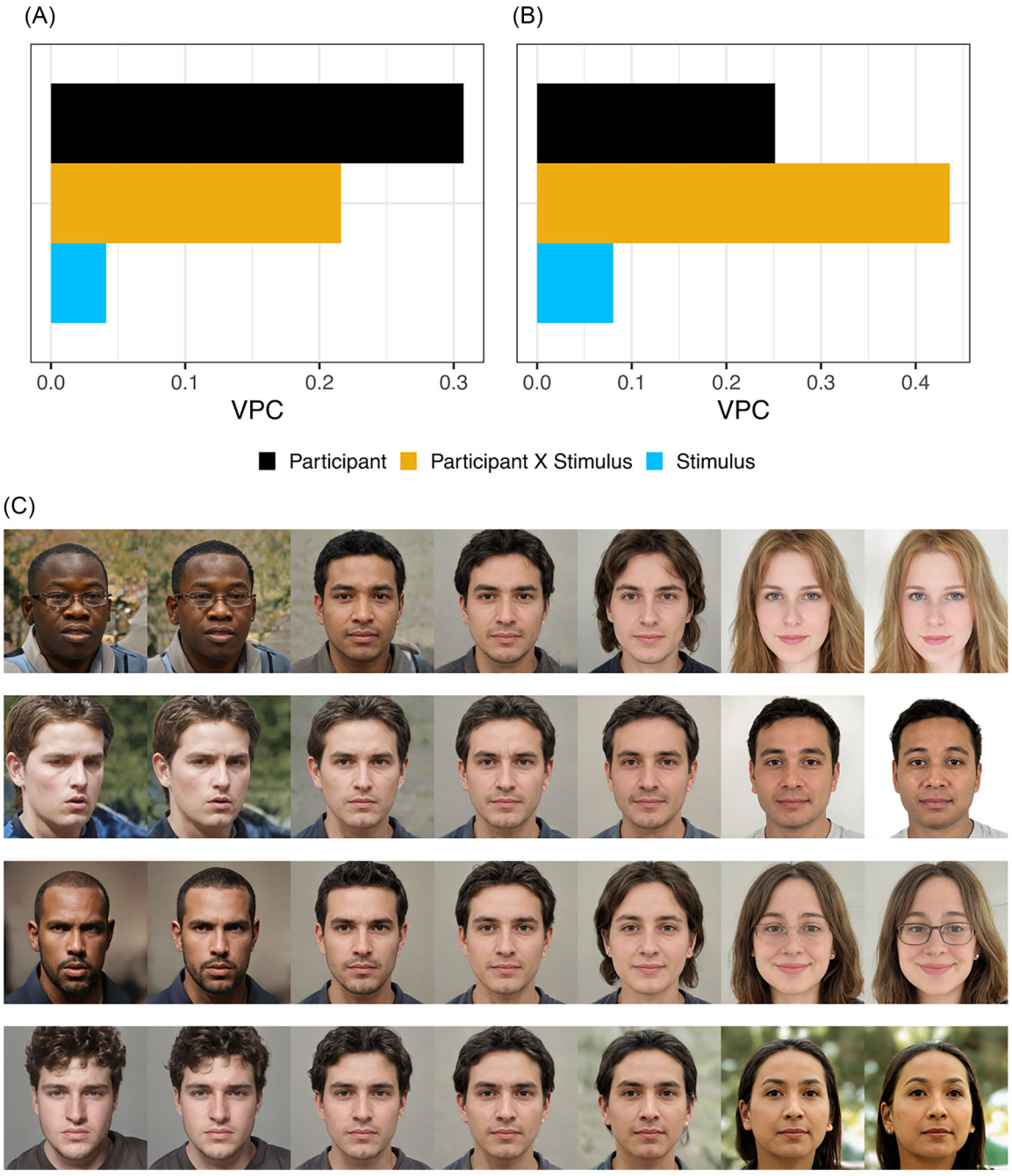

FIGURE 5 Idiosyncratic and shared contributions to trustworthiness judgments. (A) Variance partitioning coefficients (VPC) of trustworthiness judgments of neutral faces from a standardized face set.[77] Stimulus variance reflects shared contributions, whereas participant and participant × stimulus variances reflect idiosyncratic contributions (data from Albohn et al., 2024).[15] (B) VPC of trustworthiness judgments of neutral faces from a highly heterogenous face set (images collected "in the wild," varying in background, clothing, camera angle, etc.). For both sets of faces, idiosyncratic variance trumps shared variance. (C) Data-driven computational models of individuals making trustworthiness judgments. Each row represents a model fitted to the data of a single participant. As faces change from left to right, their perceived trustworthiness increases for the respective participant (adapted from Albohn et al., 2024).[78] Note the large differences between the participants' mental models of trustworthiness.

faces[81–84]—but it might be that some of the idiosyncratic differences are simply irreducible.

Nonetheless, we can build models of judgments of specific individuals, visualizing their idiosyncrasies. We can also build models of groups of individuals based on a prior theoretical interest (e.g., political affiliation[85]). Finally, the computational approach extends to any visual category of stimuli. Human judgments are highly heterogeneous and understanding those judgments would require building models that account for both shared and idiosyncratic contributions to judgments.

AUTHOR CONTRIBUTIONS

A.T. conceived of the structure of the paper and wrote the first draft. All other authors edited subsequent drafts. D.O. conducted the analyses presented in Section 1 and Figure 1, and created the model-based images of faces for Figure 2. S.U. created the model-based images of faces for Figures 3 and 4. D.N.A. created the model-based images of faces for Figure 5.

ACKNOWLEDGMENTS

This work was supported by the Richard N. Rosett Faculty Fellowship at the University of Chicago Booth School of Business.

COMPETING INTERESTS

The authors have no competing interests.

PEER REVIEW

The peer review history for this article is available at: https://publons.com/publon/10.1111/nyas.15293

REFERENCES

1. 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.
2. Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States America*, *105*, 11087–11092.
3. Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.
4. 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.
5. Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. *Advances in Experimental Social Psychology*, *63*, 189–246.
6. Brahnam, S. (2005). A computational model of the trait impressions of the face for agent perception and face synthesis. *Artificial Intelligence and Simulation of Behavior Journal*, *1*, 481–508.
7. Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*, 121–131.
8. Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*, 2261–2271.
9. Gosselin, F., & Schyns, P. G. (2003). Superstitious perceptions reveal properties of internal representations. *Psychological Science*, *15*, 505–509.
10. Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classification. *Cognitive Science*, *28*, 209–226.
11. Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision*, *2*, 105–120.
12. Todorov, A., Dotsch, R., Wigboldus, D., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass*, *5*, 775–791.
13. Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces. *IEEE*, *28*, 117–122.
14. Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 199–209.
15. Albohn, D. N., Martinez, J. E., & Todorov, A. (2024). Determinants of shared and idiosyncratic contributions to judgments of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(11), 1117–1130.
16. Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, *52*, 1428–1444.
17. Colombatto, C., Uddenberg, S., & Scholl, B. J. (2021). The efficiency of demography in face perception. *Attention, Perception, & Psychophysics*, *83*, 3104–3117.
18. Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*, 269–278.
19. Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science*, *17*, 592–598.
20. Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*, 813–833.
21. Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States America*, *104*(46), 17948–17953.
22. Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception*, *39*, 514–532.
23. Rule, N. O., Ambady, N., & Adams, R. B. (2009). Personality in perspective: Judgmental consistency across orientations of the face. *Perception*, *38*(11), 1688–1699.
24. Kurosu, A., & Todorov, A. (2017). The shape of novel objects contributes to shared impressions. *Journal of Vision*, *17*(13), 1–20.
25. Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *Journal of Personality and Social Psychology*, *111*(5), 655–664.
26. Verosky, S. C., Zoner, K. A., Marble, C. W., Sammon, M. M., & Babarinsa, C. O. (2020). Implicit responses to face trustworthiness measured with fast periodic visual stimulation. *Journal of Vision*, *20*(7), 1–12.
27. Swe, D. C., Palermo, R., Gwinn, O. S., Rhodes, G., Neumann, M., Payart, S., & Sutherland, C. A. M. (2020). An objective and reliable electrophysiological marker for implicit trustworthiness perception. *Social Cognitive and Affective Neuroscience*, *15*(3), 337–346.
28. Swe, D. C., Palermo, R., Gwinn, O. S., Bell, J., Nakanishi, A., Collova, J., & Sutherland, C. A. M. (2022). Trustworthiness perception is mandatory: Task instructions do not modulate fast periodic visual stimulation trustworthiness responses. *Journal of Vision*, *22*(11), 1–19.
29. Siddique, S., Sutherland, C. A. M., Jeffery, L., Swe, D. C., Gwinn, O. S., & Palermo, R. (2023). Children show neural sensitivity to facial trustworthiness as measured by fast periodic visual stimulation. *Neuropsychologia*, *180*, 108488.
30. Rossion, B. (2014). Understanding individual face discrimination by means of fast periodic visual stimulation. *Experimental Brain Research*, *232*, 1599–1621.
31. Cogsdill, E. J., & Banaji, M. R. (2015). Face-trait inferences show robust child–adult agreement: Evidence from three types of faces. *Journal of Experimental Social Psychology*, *60*, 150–156.
32. Cogsdill, E., Todorov, A., Spelke, E., & Banaji, M. R. (2014). Inferring character from faces: A developmental study. *Psychological Science*, *25*, 1132–1139.
33. Siddique, S., Sutherland, C. A. M., Palermo, R., Foo, Y. Z., Swe, D. C., & Jeffery, L. (2022). Development of face-based trustworthiness impressions in childhood: A systematic review and meta-analysis. *Cognitive Development*, *61*, 101131.
34. Charlesworth, T. E. S., Hudson, S.-k. T. J., Cogsdill, E. J., Spelke, E. S., & Banaji, M. R. (2019). Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology*, *55*, 1400–1413.

35. Ewing, L., Caulfield, F., Read, A., & Rhodes, G. (2015). Perceived trustworthiness of faces drives trust behaviour in children. *Developmental Science*, *18*, 327–334.
36. Ewing, L., Sutherland, C. A. M., & Willis, M. L. (2019). Children show adult-like facial appearance biases when trusting others. *Developmental Psychology*, *55*(8), 1694–1701.
37. Jessen, S., & Grossmann, T. (2016). Neural and behavioral evidence for infants' sensitivity to the trustworthiness of faces. *Journal of Cognitive Neuroscience*, *28*, 1728–1736.
38. 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.
39. Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, 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., … & Coles, N. A. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159–169.
40. Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, *127*(1), 105–118.
41. Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications*, *12*, 5168.
42. Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*, 455–460.
43. Jones, A. L., & Kramer, R. S. S. (2021). Facial first impressions form two clusters representing approach-avoidance. *Cognitive Psychology*, *126*, 101387.
44. Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*, 128–133.
45. 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.
46. Said, C., Sebe, N., & Todorov, A. (2009). Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion*, *9*, 260–264.
47. Adams, R. B., Nelson, A. J., Soto, J. A., Hess, U., & Kleck, R. E. (2012). Emotion in the neutral face: A mechanism for impression formation? *Cognition & Emotion*, *26*(3), 431–441.
48. Albohn, D. N., & Adams, R. B. Jr. (2020). Emotion residue in neutral faces: Implications for impression formation. *Social Psychological and Personality Science*, *12*(4), 479–486.
49. Albohn, D. N., & Adams, R. B. Jr. (2021). The Expressive Triad: Structure, color, and texture resemblance to emotion expressions predict impressions of neutral faces. *Frontiers in Psychology*, *12*, 612923.
50. Zebrowitz, L. A., Kikuchi, M., & Fellous, J. M. (2010). Facial resemblance to emotions: Group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology*, *98*(2), 175–189.
51. Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science*, *3*, 562–571.
52. Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., & Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, *111*(32), E3353–E3361.
53. Jaeger, B., & Jones, A. L. (2022). Which facial features are central in impression formation? *Social Psychological and Personality Science*, *13*(2), 553–561.
54. 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 of the United States America*, *119*(17), E2115228119.
55. 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.
56. Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. *Advances in Experimental Social Psychology*, *30*, 93–161.
57. Zebrowitz, L. A. (1997). *Reading faces: Window to the soul?*, Boulder, CO: Westview Press.
58. Oh, D., Wedel, N., Labbree, B., & Todorov, A. (2023). Trustworthiness judgments without the halo effect: A data-driven computational modeling approach. *Perception*, *52*, 590–607.
59. Oh, D., Dotsch, R., & Todorov, A. (2019). Contributions of shape and reflectance information to social judgments from faces. *Vision Research*, *165*, 131–142.
60. Torrance, J. S., Wincenciak, J., Hahn, A. C., DeBruine, L. M., & Jones, B. C. (2014). The relative contributions of facial shape and surface information to perceptions of attractiveness and dominance. *PLoS ONE*, *9*, e104415.
61. Jablonski, N. G., & Chaplin, G. (2000). The evolution of human skin coloration. *Journal of Human Evolution*, *39*, 57–106.
62. Said, C. P., & Todorov, A. (2011). A statistical model of facial attractiveness. *Psychological Science*, *22*, 1183–1190.
63. Funk, F., Walker, M., & Todorov, A. (2017). Modelling perceptions of criminality and remorse from faces using a data-driven computational approach. *Cognition and Emotion*, *31*, 1431–1443.
64. Toscano, H., Schubert, T. W., Dotsch, R., Falvello, V., & Todorov, A. (2016). Physical strength as a cue to dominance: A data-driven approach. *Personality and Social Psychology Bulletin*, *42*, 1603–1616.
65. Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions from faces. *Psychological Science*, *30*, 65–79.
66. Abir, Y., Sklar, A., Dotsch, R. R., Todorov, A., & Hassin, R. (2018). The determinants of consciousness of human faces. *Nature Human Behaviour*, *2*, 194–199.
67. Cao, R., Li, X., Todorov, A., & Wang, S. (2020). A flexible neural representation of faces in the human brain. *Cerebral Cortex Communications*, *1*, 1–12.
68. Jaeger, B., Todorov, A., Evans, A. M., & van Beest, I. (2020). Can we reduce facial biases? Persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, *90*, 10404.
69. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). *Analyzing and improving the image quality of StyleGAN*. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8107–8116). IEEE.
70. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, *34*, 852–863.
71. 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.
72. Albohn, D. N., Uddenberg, S., & Todorov, A. (2022). A data-driven, hyper-realistic method for visualizing individual mental representations of faces. *Frontiers in Psychology*, *13*, 997498.
73. Bjornsdottir, R. T., Hehman, E., & Human, L. J. (2021). Consensus enables accurate social judgments. *Social Psychological and Personality Science*, *13*(6), 1010–1021.
74. Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology*, *113*(4), 513–529.
75. Lavan, N., & Sutherland, C. A. (2024). Idiosyncratic and shared contributions shape impressions from voices and faces. *Cognition*, *251*, 105881.

76. Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology*, *28*, 333–361.
77. Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*, 1122–1135.
78. Albohn, D. N., Uddenberg, S., & Todorov, A. (2025). Individualized models of social judgments and context-dependent representations. *Scientific Reports*, *15*, 4208. https://doi.org/10.1038/s41598-025-86056-1
79. Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A. M., Smoller, J. W., Kwok, H., Samuel E. Anthony, Ken Nakayama, Gillian Rhodes, & Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology*, *25*, 2684–2689.
80. Sutherland, C. A. M., Burton, N. S., Wilmer, J. B, Blokland, G. A. M., Germine, L., Palermo, R., Collova, J. R., & Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences of the United States America*, *117*, 10218–10224.
81. Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour*, *1*, 1–6.
82. Sofer, C., Dotsch, R., Oikawa, M., Oikawa, H., Wigboldus, D. H. J., & Todorov, A. (2017). For your local eyes only: Culture-specific face typicality influences perceptions of trustworthiness. *Perception*, *46*, 914–928.
83. Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*, 779–785.
84. Verosky, S. C., & Todorov, A. (2013). When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. *Journal of Experimental Social Psychology*, *49*, 661–669.
85. Uddenberg, S., Nguyen, Z., Albohn, D. N., & Todorov, A. *Hyper-realistic reverse correlation reveals a gender bias in representations of leadership*. (Under review).

How to cite this article: Todorov, A., Oh, D. W., Uddenberg, S., & Albohn, D. N. (2025). Face evaluation: Findings, methods, and challenges. *Ann NY Acad Sci.*, *1545*, 28–37. https://doi.org/10.1111/nyas.15293