AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2025, Vol. 154, No. 3, 644–657
https://doi.org/10.1037/xge0001717

Goals Bias Face Perception

Yi-Fei Hu[1], Joseph Heffner[1, 2], Apoorva Bhandari[1], and Oriel FeldmanHall[1, 3]
[1] Department of Cognitive and Psychological Sciences, Brown University
[2] Department of Psychology, Yale University
[3] Carney Institute for Brain Science, Brown University

Faces—the most common and complex stimuli in our daily lives—contain multidimensional information used to infer social attributes that guide consequential behaviors, such as deciding who to trust. Decades of research illustrates that perceptual information from faces is processed holistically. An open question, however, is whether goals might impact this perceptual process, influencing the encoding and representation of the complex social information embedded in faces. If an individual were able to factorize information so that each dimension is separately represented, it might enable flexibility. Having a goal, for example, might mean that only goal-relevant dimensions are leveraged to inform behavior. Whether people are able to build such factorized representations remains unknown, largely due to natural correlations between social attributes. We overcome these confounds using a new statistical face model that orthogonalizes perceived facial attractiveness and trustworthiness. Across three experiments (N = 249), we observe that only in some contexts can humans successfully factorize multidimensional social information. When there is a clear goal of assessing another's trustworthiness, people successfully decompose these social attributes. The more an individual factorizes, the more they entrust money to others in a subsequent trust game. However, when the goal is to assess attractiveness, irrelevant information about trustworthiness is so potent that it biases how attractive someone is perceived—a trustworthiness "halo effect." In contrast, in goal-agnostic environments, we do not find any evidence of factorization; instead, people encode multidimensional social information in an entwined and holistic fashion that distorts their perceptions of social attributes.

Public Significance Statement

Humans swiftly extract social information from faces. Despite decades of research suggesting that people process faces in a holistic manner, it remains unknown whether humans can parsimoniously extract only goal-relevant information to achieve a particular goal (e.g., asking people who look trustworthy for directions while ignoring other social attributes). Leveraging a newly developed statistical face model that disentangles the high correlation between attractiveness and trustworthiness, we observe that when there is a clear goal, people only rely on goal-relevant information to make decisions while ignoring goal-irrelevant information. However, this effect depends on the goal. In some cases, perceptual information about another's trustworthiness is so potent that it biases people's judgments about how attractive one appears, suggesting a trustworthiness "halo effect." In goal-agnostic environments, people are unable to ignore goal-irrelevant information and instead encode social information in a holistic manner. These results reflect the flexible nature of social perception.

Keywords: face perception, factorization, gestalt, halo effect, trustworthiness

Supplemental materials: https://doi.org/10.1037/xge0001717.supp

This article was published Online First January 13, 2025.

Wilma Bainbridge served as action editor.

Oriel FeldmanHall https://orcid.org/0000-0002-0726-3861

The materials and raw data for all experiments are available via the Open Science Framework at https://osf.io/5awbs/. Results in this work have not been presented at any conferences.

The authors have no conflicts of interest to disclose. This work is supported by the Carney Innovation Grant from the Robert J. and Nancy D. Carney Institute for Brain Science (Oriel FeldmanHall) and the National Science Foundation award 2123469 (Oriel FeldmanHall and Apoorva Bhandari). The authors thank Logan Bickel for helping with data collection, Brandon P. Labbree and DongWon Oh for helping with face model construction, and David Badre for suggestions on the project.

Conceptualization: Yi-Fei Hu, Joseph Heffner, Apoorva Bhandari, and Oriel FeldmanHall. Formal analysis: Yi-Fei Hu and Joseph Heffner. Funding acquisition: Oriel FeldmanHall and Apoorva Bhandari. Investigation: Yi-Fei Hu and Joseph Heffner. Methodology: Yi-Fei Hu, Joseph Heffner, Apoorva Bhandari, and Oriel FeldmanHall. Supervision: Oriel FeldmanHall and Apoorva Bhandari. Writing: Yi-Fei Hu, Joseph Heffner, Apoorva Bhandari, and Oriel FeldmanHall.

Yi-Fei Hu played a lead role in formal analysis and visualization, and an equal role in conceptualization, data curation, methodology, writing–original draft, and writing–review and editing. Joseph Heffner played a supporting role in formal analysis, writing and editing, and an

continued

The world is full of information. Some of it can be used to reduce the uncertainty permeating the environment. However, not all available information is useful, relevant, or necessary when pursuing a particular goal. Imagine, for example, you are lost at a busy intersection in a new city. With just a glimpse of those passing by, you can swiftly form an impression of each stranger's age, attractiveness, trustworthiness, and socioeconomic status (Asch, 1946; Bar et al., 2006). Faces are a rich source of information spanning a multiplicity of dimensions (Jack & Schyns, 2017; Todorov et al., 2015). Indeed, the information gleaned from faces can be used to make rapid assessments about a person's social category, such as their gender or race (Cloutier et al., 2005; Quinn & Macrae, 2005), political competence (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Olivola & Todorov, 2010; Todorov et al., 2005), trustworthiness (Rule et al., 2013; Todorov et al., 2009; van 't Wout & Sanfey, 2008; Willis & Todorov, 2006), or socioeconomic class (Bjornsdottir et al., 2024; Bjornsdottir & Rule, 2020) in as little as half a second. If we consider our goal of getting directions, only the information about trustworthiness is relevant, while age, attractiveness, and socioeconomic status are far less so. How one's goals influence the encoding of information is an open question—one that concerns attention, perception, and memory researchers.

One possible way in which goals might shape how complex social information is encoded is through decomposition (Hyvärinen & Oja, 2000; Kingma & Welling, 2022; Wold et al., 1987). With decomposition, social information is broken down into separate dimensions (Bengio et al., 2014). Decomposed information can be encoded in a factorized representation, where each dimension is stored separately, or a filtering mechanism can gate out any factorized information that is irrelevant to the goal. An alternative possibility is that people encode multidimensional information to form a gestalt representation (Koffka, 1922; Wagemans et al., 2012), one that can be used to service multiple different goals as they become pertinent over time. Here we investigate how goals shape the encoding of social information for adaptive behavior.

Evidence from machine learning and artificial intelligence favors the first account: social information from faces is read out in a factorized manner (Lee & Seung, 1999; Liu et al., 2015; Parkhi et al., 2015; Soulos & Isik, 2024; Turk & Pentland, 1991), where social attributions (e.g., trustworthiness, attractiveness) are combinations of decomposed facial features (i.e., eye size, skin luminesce, etc.; Jaeger & Jones, 2022; Oosterhof & Todorov, 2008; Peterson et al., 2022; Vernon et al., 2014). This mapping between social attributions and decomposed facial features allows for maximal flexibility. Machines can simply extract any social attribution or any decomposed facial features independently from one another, disentangling the representation to improve learning success (Bengio et al., 2014; Lake et al., 2017). Machines can also independently tune any subset (or all) of the physical features (i.e., nose size, skin luminesce, etc.) to manipulate a particular social attribution—such as attractiveness. This is a technique often adopted by social perception researchers, and is akin to tuning neurons in the input layer in a two-layer network to generate different outcomes in the output layer (He et al., 2019; Karras et al., 2021; Kulkarni et al., 2015). While these data-driven methods are fast and efficient in creating infinite faces with different social attributions by adjusting any number of relevant physical features, research on how humans encode social information from faces illustrates a largely different process.

Instead, people seem to rapidly extract social attributions from faces in a holistic, gestalt manner. Such holistic processing manifests in two ways: First, faces are perceived as unified wholes, and second, social attributions, which are read out from faces, appear to influence one another, such that they are represented in an intertwined manner. Composite-face paradigms (Young et al., 1987), which serve as the bread-and-butter methodology for interrogating face processing, demonstrate that people mainly rely on holistic—rather than decomposed—processing to make social impressions (Farah et al., 1998; Maurer et al., 2002; Todorov et al., 2010). As just one example, when a top portion of an attractive face is combined with a bottom portion of an unattractive face, the attractive top portion of the composite-face is perceived as less attractive, suggesting that faces are processed as a whole (Abbas & Duchaine, 2008)—an effect that extends to trust (Todorov et al., 2010), gender (Baudouin & Humphreys, 2006), race (Michel et al., 2007), and emotional expressions (Calder et al., 2000). Furthermore, inferred social attributions—for example, trustworthiness and attractiveness—from faces are often correlated with one another, further hinting that social dimensions are processed in an entwined manner (Jones et al., 2021; Oosterhof & Todorov, 2008). Additional evidence that social judgments are not decomposed but rather processed in a gestalt manner comes from the well-known "halo effect," where a physically attractive face is associated with other positive social features, such as greater competence or intelligence (Dion et al., 1972; Eagly et al., 1991; Langlois et al., 2000).

Despite decades of work illustrating that humans holistically process multidimensional facial information, advances in artificial intelligence tout factorization as a powerful mechanism that can efficiently encode information for adaptive decisions (Bengio et al., 2014; Chen et al., 2016; Higgins et al., 2017; Lake et al., 2017). However, the natural confounds between attributes in face perception (Jones et al., 2021; Oosterhof & Todorov, 2008), combined with a historical legacy favoring goal-invariant paradigms, have left unanswered questions of whether people are capable of factorizing multidimensional information from faces in the service of a particular goal. Even though people can extract a single social attribution when making decisions with a fixed goal in mind (e.g., being asked to judge someone's competence; Antonakis & Dalgas, 2009; Todorov et al., 2005), if other goal-irrelevant social attributions (e.g., attractiveness) are

equal role in conceptualization, data curation, and methodology. Apoorva Bhandari played a supporting role in funding acquisition, supervision, writing–original draft, and writing–review and editing and an equal role in conceptualization and methodology. Oriel FeldmanHall played a lead role in funding acquisition and supervision and an equal role in conceptualization, methodology, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Oriel FeldmanHall, Department of Cognitive and Psychological Sciences, Brown University, 190 Thayer Street, Providence, RI 02912, United States. Email: oriel.feldmanhall@brown.edu

present alongside goal-relevant attributions (e.g., competence), it is almost impossible to decipher whether the representation built for pursuing a particular goal factorizes social information, or, whether multiple pieces of information have been encoded into a holistic, integrated representation. Evidence of factorization would require that following encoding, subsequent decisions remain unaffected by goal-irrelevant information. This parsimonious solution is akin to encoding a stranger's trustworthiness at the busy intersection when asking for directions without entangling any other socially irrelevant attributions (e.g., attractiveness, age).

There are of course times when there is no clear goal at hand. How is multidimensional information encoded in the absence of a goal? A solution is to encode all available information and selectively invoke relevant dimensions once a decision needs to be executed down the line. In these cases, an additional question arises about the representations maintained in goal-agnostic environments. Are these representations created from information that has been factorized (in which case, information along some dimensions can be recalled without interference from other dimensions), or, from information integrated in such a way that the representation is truly gestalt? As applied to our example above, when there is no goal, an individual may extract multiple social attributions and then either build a representation in which each dimension is separately stored, or, where all dimensions are interwoven and entangled in a holistic manner.

To test these open questions about how goals impact the encoding and representation of multidimensional social information, we leverage a working memory gating paradigm (Chatham et al., 2014) that enables an assessment of whether social information can be decomposed or factorized, and if so, whether any information is subsequently filtered out in service of meeting the current goal. We present people with unfamiliar faces in different contexts, in some cases, particular goals are known to participants, and in others, there is no obvious goal. We probe the nature of each representation, hypothesizing that having a specific goal is associated with decomposed, factorized representations while goal agnostic environments obligatorily lead to integrating information into a holistic representation. To test the behavioral relevance yoked to each type of representation, we then asked participants to play a trust game with these individuals (Berg et al., 1995), which enabled us to interrogate how each representational format impacts subsequent social choices.

Experiment 1: Orthogonalizing Social Information

Considering their ubiquitous impact on our daily social decisions (Maestripieri et al., 2017; Todorov, 2008), we selected the dimensions of attractiveness and trustworthiness as the two social attributes to interrogate. To investigate the cases in which people factorize versus maintain a gestalt representation, we need a set of faces in which social attributions are completely orthogonal. That is, only one social dimension should be relevant to a particular goal, while providing zero information about another goal. Given that attractiveness and trustworthiness are typically highly correlated in real human faces (Jones et al., 2021; Oh et al., 2023; Todorov, 2008), we first needed to create a face set in which these two social dimensions were independent from one another. We therefore built a statistical face model that generated a stimulus set of 150 face images in which the perceived attractiveness and trustworthiness of the faces are algorithmically orthogonal. We then validated the orthogonality between attractiveness and trustworthiness in our stimulus set using participant responses.

Method

A Statistical Face Model Orthogonalizing Attractiveness and Trustworthiness

Our statistical face model is derivative of two existing data-driven face models: an attractiveness and trustworthiness face model (Todorov et al., 2013). These two models leverage stimuli from FaceGen (Singular Inversions, Toronto, Canada). FaceGen defines the state space of human faces, where each face is represented as a vector with 100 parameters in 100 facial principal components (PC; 50 shape PCs and 50 texture PCs). A change along any principle component leads to a holistic change in appearance, which is orthogonal to changes in other PCs (Blanz & Vetter, 1999). In the attractiveness and trustworthiness models, 100 facial PCs are mapped onto the perceived attractiveness or trustworthiness, based on human ratings. In other words, the perceived attractiveness and trustworthiness of a face are modeled as a weighted linear combination of the 100 facial PCs (Eqs. 1 and 2). With these two models, we can freely manipulate the perceived attractiveness and trustworthiness of a face by adjusting any of the 100 facial PCs. However, due to the correlational nature between attractiveness and trustworthiness from human judgments, manipulating a face's perceived attractiveness automatically changes its perceived trustworthiness and vice versa. That is, the 100 weights of facial PCs in the attractiveness model are highly correlated with the 100 weights in the trustworthiness model. Thus, to create a face model in which the perceived attractiveness and trustworthiness can be manipulated independently without affecting the other dimension, we applied the same methodology described in Oh et al. (2019, 2023) to regress out trustworthiness-related information from the attractiveness model and attractiveness-related information from the trustworthiness model.

$$\text{Attractiveness} = \sum_{i=1}^{100} \beta_{Ai} \times \text{PC}_i, \tag{1}$$

$$\text{Trustworthiness} = \sum_{i=1}^{100} \beta_{Ti} \times \text{PC}_i, \tag{2}$$

We applied a linear regression that predicts the 100 weights of facial PCs in the attractiveness model using the 100 weights in the trustworthiness model (Eq. 3) and retained the residuals of this linear regression as the new 100 weights of the facial PCs to create an Attractiveness $\perp$ Trustworthiness model (Eq. 5). In this Attractiveness $\perp$ Trustworthiness model, adjusting any facial PCs would only lead to a change in the perceived attractiveness of a face while holding its perceived trustworthiness constant. We applied the same logic in reverse to create a Trustworthiness $\perp$ Attractiveness model (Equations 4 and 6). This gives us a statistical face model that allows us to manipulate either the perceived attractiveness or trustworthiness of a face, without interference from the other dimensions.

$$\beta_A = k_{A\propto T} \times \beta_T + \varepsilon_{A\perp T}, \quad (3)$$

$$\beta_T = k_{T\propto A} \times \beta_A + \varepsilon_{T\perp A}, \quad (4)$$

$$\text{Attractiveness} \perp \text{Trustworthiness} = \sum_{i=1}^{100} \varepsilon_{A\perp Ti} \times \text{PC}_i, \quad (5)$$

$$\text{Trustworthiness} \perp \text{Attractiveness} = \sum_{i=1}^{100} \varepsilon_{T\perp Ai} \times \text{PC}_i. \quad (6)$$

Final Face Stimulus Set

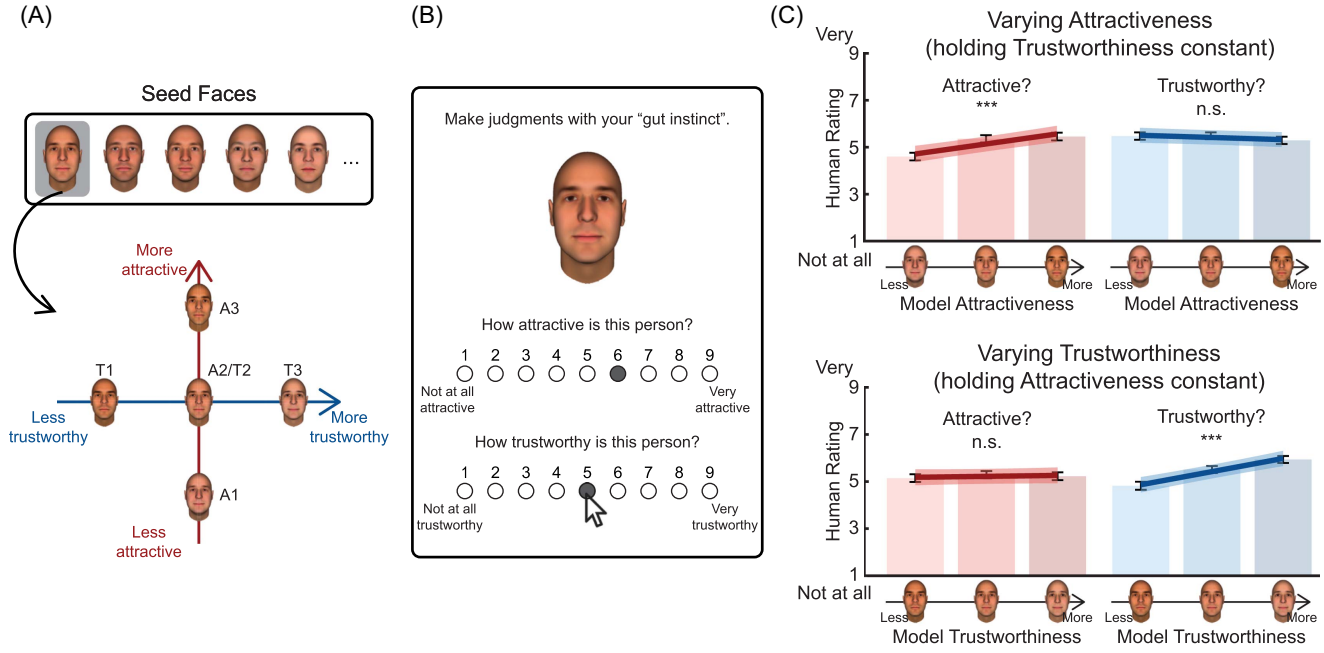
To create a face stimulus set, we first generated 25 seed faces with different identities which were sufficiently distinguishable. We then adjusted the 100 facial PCs of the 25 seed faces to project them at 0 *SD* on the Attractiveness ⊥ Trustworthiness dimension, rendering them all neutrally attractive. We then projected each of the 25 neutrally attractive faces at −2.63, 0, and +2.63 *SD* on the Trustworthiness ⊥ Attractiveness dimension to create a set of untrustworthy, neutrally trustworthy, and trustworthy faces with neutral levels of attractiveness for each identity. We did the same thing for the other dimension, which together generated a set of six faces with different levels of perceived attractiveness and trustworthiness (Figure 1A), resulting in a final stimulus set of 150 face images.

To validate the face set, we recruited 115 participants (38 participants reported their genders as female, 76 as male, and 1 as nonbinary; average age = 33.24 ± 9.24) from Amazon Mechnical Turk and asked them to rate the perceived attractiveness and trustworthiness of each face randomly presented one after another on two 9-point Likert scales (Figure 1B). The response time of each face was unlimited. Each participant first rated all 150 face images once. Then each participant rated a randomly selected subset of 25 face images for a second time as an intra-rater reliability measure. Fifteen participants were removed from the final analyses because of their low intra-rater reliability (test–retest correlation smaller than 0). One additional participant was excluded due to their highly repetitive responses (more than 80% of their responses were exactly the same and the default response). This yielded a final sample size of 99 participants. The inclusion of these 16 participants does not change the results.

Transparency and Openness

All experiments were approved by Brown University's internal review board, and informed consent was obtained by all participants before engaging in any of the studies reported here. The experiments reported in this article were not formally preregistered. The face stimulus set and raw data for all experiments are available via the Open Science Framework at https://osf.io/5awbs/.

Figure 1
Experiment 1: Design and Results



Note. (A) A set of six faces generated from each of the 25 seed faces with a statistical face model orthogonalizing attractiveness and trustworthiness. (B) Experimental display for the social attribution rating task. (C) Results of the face model estimation validation. The *x*-axis reflects the model's attribute (less [attractive/trustworthy] = −2.63 *SD*, neutral = 0 *SD*, more = 2.63 *SD*); the *y*-axis shows participants' average ratings on attractiveness (red) and trustworthiness (blue). Bars reflect average ratings across participants and faces. Curves reflect participants' attribution rating predicted by model estimates. A = attractive; T = trustworthy; n.s. = not significant. See the online article for the color version of this figure.
*** $p < .001$.

Results

Interrater agreements in attractiveness and trustworthiness of the model faces are both high with an interclass correlation coefficient (ICC[2, k], i.e., absolute agreement across all raters; Koo & Li, 2016) of 0.774 and 0.816 and a Cronbach's α of 0.896 and 0.905, respectively. Participants rated faces that had higher model estimates of attractiveness as more attractive ($\beta = 0.16 \pm 0.03$, $t = 6.05$, $p < 10^{-8}$), but not more trustworthy ($\beta = -0.03 \pm 0.02$, $t = -1.92$, $p = .06$). Faces with higher model estimates of trustworthiness were rated as more trustworthy ($\beta = 0.21 \pm 0.03$, $t = 7.42$, $p < 10^{-12}$), but not more attractive ($\beta = 0.02 \pm 0.02$, $t = 0.66$, $p = .51$). In other words, our orthogonal model successfully disentangled the notoriously high positive correlations between perceived attractiveness and trustworthiness of faces (Jones et al., 2021; Oh et al., 2023; Oosterhof & Todorov, 2008). We also successfully orthogonalized attractiveness and trustworthiness in an additional study (Experiment 1b) in which participants were asked to make predictions about how attractive and trustworthy they thought other people would find these faces. Results replicate Experiment 1 (see Supplemental Material for more details).

Experiment 2: Do People Only Encode Relevant Social Information Given a Specific Goal?

When working toward a specific goal, do people factorize and parsimoniously encode only the relevant social information from faces, while filtering out the irrelevant information? To answer this question, we leveraged a working memory input gating paradigm (Chatham et al., 2014), where participants were informed of a specific goal on each trial before being presented with a picture of a person's face. In some trials, participants were informed that they would be making predictions about a person's attractiveness; while in other trials, they would make predictions about a person's trustworthiness. In addition, participants played a trust game with these people, which allowed us to interrogate how decomposed (or gestalt) representations influence social choices.

Method

Participants

A total of 84 in-lab participants (48 participants reported their genders as female, and 36 as male; average age = 19.60 ± 1.75) from Brown University community were recruited and either compensated for their time with a monetary reward or course credit. One participant was removed from analyses due to their high timeout rates, yielding a final sample size of 83. The inclusion of this participant's data does not change the findings.

Design

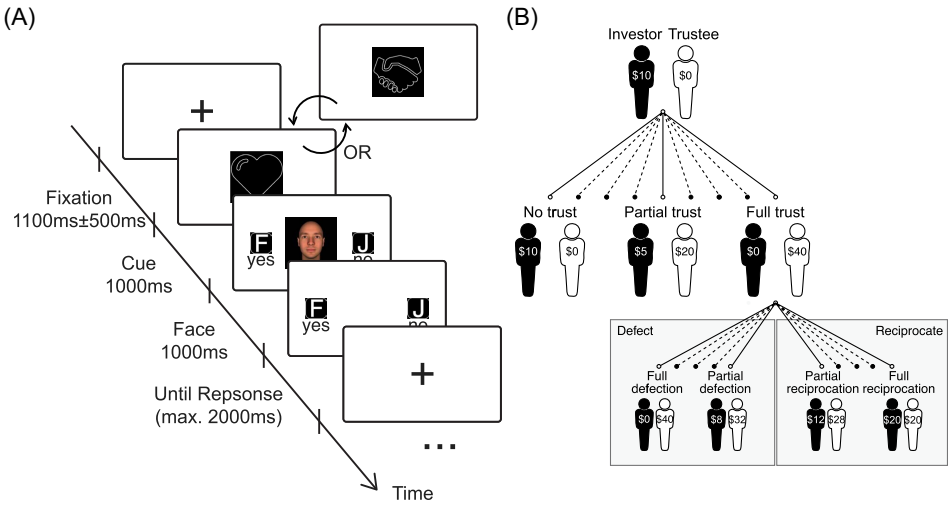
In the first task, we leveraged a working memory input gating paradigm, where the goal is revealed to participants at the start of the trial. Participants were told that people from a previous study had rated a set of faces on two dimensions: attractiveness and trustworthiness; and the job of the current participant was to predict whether each face was either considered above- or below-average attractiveness/trustworthiness by the previous participants. Validated faces ($N = 150$) from Experiment 1 were used as stimuli. From this initial validated set of faces, 15 sets of faces (each set comprised of six faces) were selected to be presented to participants. Thus, participants were presented with 90 faces over the course of the first task. Each face was presented twice; the participant was asked to judge attractiveness/trustworthiness on separate trials. On each trial, participants were first presented with a cue denoting the goal of the trial: to judge attractiveness, denoted by a heart icon, or, to judge trustworthiness, denoted by a handshake icon. The use of simple icons to represent the goals of trials matched with the original working memory gating paradigm (Chatham et al., 2014). After the cue, participants were then presented with a face for 1,000 ms. Participants responded in a binary manner (yes = above average/no = below average) by pressing one of two keys. The mapping between the keys and the responses (yes/no) switched across trials to prevent participants from performing habitual key-pressing behaviors. No feedback was provided. Participants were instructed to respond within 3 s. If participants did not make a response within 3 s, a warning in red read "Respond Faster!" would pop up on the screen. Between trials, there was a 1,500- to 2,000-ms trial interval. See Figure 2A for an illustration of an experimental trial. Before the formal experiment, participants completed five practice trials to become familiar with the task, especially, the use of icon cues.

In the second task, participants played a trust game (Berg et al., 1995) with the remaining validated faces (10 sets of 25 were selected, for a total of 60 faces) that were not used in the first task. Each face was presented once. On each trial, participants were endowed with $10 and acted as the investor, deciding whether to entrust their money with the other player (Figure 2B). Participants were informed that any money invested would be multiplied four times, and the other player could then either share the money back with the participant (reciprocate) or keep the money for themselves (defect). Participants made only one decision with each person and to prevent learning biases over the task, participants never received feedback regarding whether the other player reciprocated or defected.

Results

If people are capable of decomposing information from faces and representing it in a factorized manner, we should expect to see participants' performance predicted only by the relevant information, while remaining unaffected by the irrelevant information. For example, when participants are asked to make predictions about a face's attractiveness, they may encode information pertaining to attractiveness and not trustworthiness, while exhibiting this pattern in reverse when asked to make judgments about trustworthiness. We leveraged mixed-effects logistic regressions (maximal models without interaction terms between model estimates of face attractiveness and trustworthiness) to test participants' predictions about each perceived social attribute, which revealed that when asked to make predictions about the attractiveness of faces, participants were able to extract the relevant attractive information ($\beta_{judgeA\_A} = 0.412 \pm 0.038$, $t = 10.788$, $p < 10^{-26}$; Figure 3B); however, they were incapable of filtering out trustworthy information, which influenced their judgments along the attractiveness dimension ($\beta_{judgeA\_T} = 0.157 \pm 0.037$, $t = 4.233$, $p < 10^{-4}$; Figure 3C). That is, the more a face was judged to be trustworthy, the more it was also perceived as attractive—despite the fact that these faces had the same level of model estimated attractiveness. Prior work

Figure 2
Experiment 2: Design



Note. (A) Experimental procedure illustrating one trial in social attribute prediction task adopting the working memory input gating paradigm in Experiment 2. Participants were given a clear goal of either judging the attractiveness or trustworthiness before a face was presented to them. (B) The trust game. Participants were endowed with $10 at the beginning of each round. Any money participants decided to entrust is multiplied four times, and the trustee then can decide to return any portion of the money from full defection (0%) to full reciprocation (50%). See the online article for the color version of this figure.

suggests that when attractiveness of a face is controlled for, the more trustworthy it looks, the more positive emotion it shows (Oh et al., 2023). Could this trustworthy halo effect be explained by increasingly positive emotional expressions as trustworthiness increases? To rule out this possibility, we extracted positive emotional facial features, treated them as control variables, and reran the analysis (see Supplemental Material for details). We find no evidence that the trustworthy halo effect is explained by a perception of positive emotion expressions from the model faces. That is, the same outcomes are observed when controlling for positive emotional features: When asked to predict attractiveness, participants relied on attractive information ($\beta_{judgeA\_A} = 0.384 \pm 0.040$, $t = 9.593$, $p < 10^{-20}$; Supplemental Figure S4A) but failed to filter out trustworthy information ($\beta_{judgeA\_T} = 0.178 \pm 0.043$, $t = 4.132$, $p < 10^{-4}$; Supplemental Figure S4B). Interestingly, such a halo effect was only observed for information pertaining to trustworthiness. When participants made predictions about the trustworthiness of faces, they reliably extracted trustworthy information from faces ($\beta_{judgeT\_T} = 0.512 \pm 0.051$, $t = 9.995$, $p < 10^{-22}$; Figure 3E), while successfully decomposing and then filtering out attractiveness information—thereby preventing attractiveness information from affecting their judgments about trustworthiness ($\beta_{judgeT\_A} = 0.038 \pm 0.035$, $t = 1.090$, $p = .276$; Figure 3D).

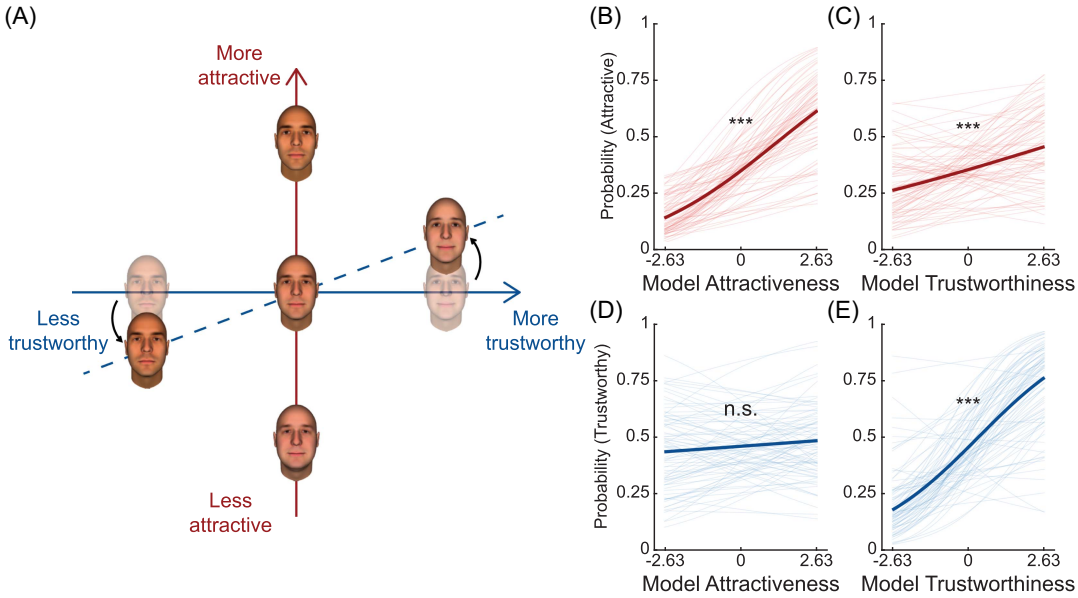
How did decomposition and filtering out irrelevant information about a person's attractiveness bias decisions to trust? We leveraged mixed-effects linear regressions to test how much money participants entrust to faces, given differing levels of attractiveness and trustworthiness. First, participants only relied on trustworthiness information to inform their decisions to trust ($\beta_{modelT} = 0.456 \pm 0.045$, $t = 10.050$, $p < 10^{-22}$; Figure 4A blue line) while largely filtering out information about attractiveness ($\beta_{modelA} = 0.065 \pm 0.034$, $t = 1.910$, $p = .056$; Figure 4A red line). We then probed how much an individual was able to factorize social attributes influenced their decisions to trust. We leveraged linear regressions that included interaction terms between model estimates of face attributes (attractiveness and trustworthiness), and the individual-level beta values extracted from the previous logistic regression (which represents the degree to which participants rely on relevant facial information make social attribute judgments, e.g., trustworthiness information for trustworthiness judgment). Results reveal that the extent to which participants relied on trustworthiness information to inform their decisions to trust was modulated by how much they factorized trustworthiness information ($\beta_{modelT \times \beta_{judgeT\_T}} = 0.733 \pm 0.115$, $t = 6.377$, $p < 10^{-7}$; Figure 4B). In contrast, how much participants relied on attractiveness information to judge a person's attractiveness in the memory gating paradigm did not predict money trusted ($\beta_{modelT \times \beta_{judgeA\_A}} = -0.012 \pm 0.184$, $t = -0.068$, $p = .946$; Figure 4B).

In short, when given a specific goal, participants could only sometimes filter out irrelevant information: Participants failed to completely gate trustworthy information, which ultimately distorted their judgments of a person's attractiveness. However, they were able to successfully prevent information about attractiveness from influencing their judgments about another's trustworthiness. Similarly, participants were able to exclusively rely on trustworthy information to guide their decisions to entrust money to others. The degree to which they relied on this trustworthy information was modulated by their ability to successfully factorize this information. But what mechanism is driving this asymmetric halo effect?

One possibility is that trustworthiness is so potent that this social information cannot be filtered out, even when an input gate is

Figure 3

Experiment 2: Gating Task Results

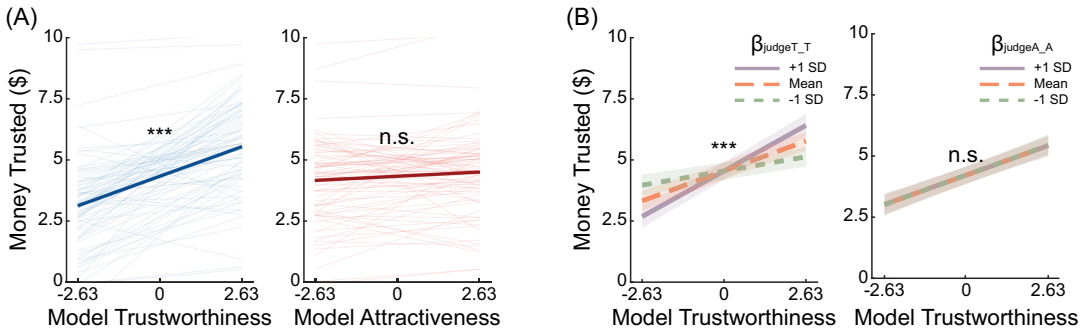


Note. (A) A toy illustration of how social dimensions become distorted when trustworthy information is not filtered. More trustworthy faces are perceived to be more attractive, but this relationship was only observed in this direction and not in the reverse. (B–E) Participants' judgments on social dimensions predicted by relevant and irrelevant information. The *x*-axis illustrates the model estimates of social attributes (attractiveness and trustworthiness) of the faces. The *y*-axis denotes the average probability of participants judging the faces as above average attractiveness (red) and above average trustworthiness (blue). n.s. = not significant. See the online article for the color version of this figure.
*** $p < .001$.

applied; rather trustworthiness information is irresistibly encoded, thereby affecting representations of other visual social attributes. In contrast, other less potent social information, such as attractiveness, can be effectively filtered out. If this is the case, it would suggest that only some information can be filtered in service of a goal. It is also possible, however, that no filtering mechanism was applied at all, and that instead, information from both dimensions was encoded, such that perceived trustworthiness distorted the representation of perceived attractiveness. If this were the case, it would suggest that only some attributes can be represented in a

Figure 4

Experiment 2: Trust Game Results



Note. (A) Money entrusted as a function of the model estimated trustworthiness (blue) and attractiveness (red). The *x*-axis reflects model estimates of each social attribute, while the *y*-axis is the average amount entrusted in the trust game. (B) Money entrusted, which is predicted by the model estimated trustworthiness (Panel A), is modulated by the degree to which participants relied on the relevant information (trustworthiness). n.s. = not significant. See the online article for the color version of this figure.
*** $p < .001$.

factorized manner, while others cannot. To test these competing hypotheses, we conducted a third experiment, leveraging an output gating paradigm where the goal was not made clear to participants until after social information was provided, essentially rendering any filtering mechanism useless.

Experiment 3: Maintaining Factorized Representations of Social Attributions

An output gating paradigm, where the information is provided *prior* to a goal, helps to answer a more fundamental question about representation. If multiple social dimensions are extracted from faces and encoded without an input filter, do people represent multidimensional information in a factorized manner? Evidence of factorization would be to make judgments about the relevant dimension without interference from the irrelevant dimension once the goal becomes clear.

Method

Participants

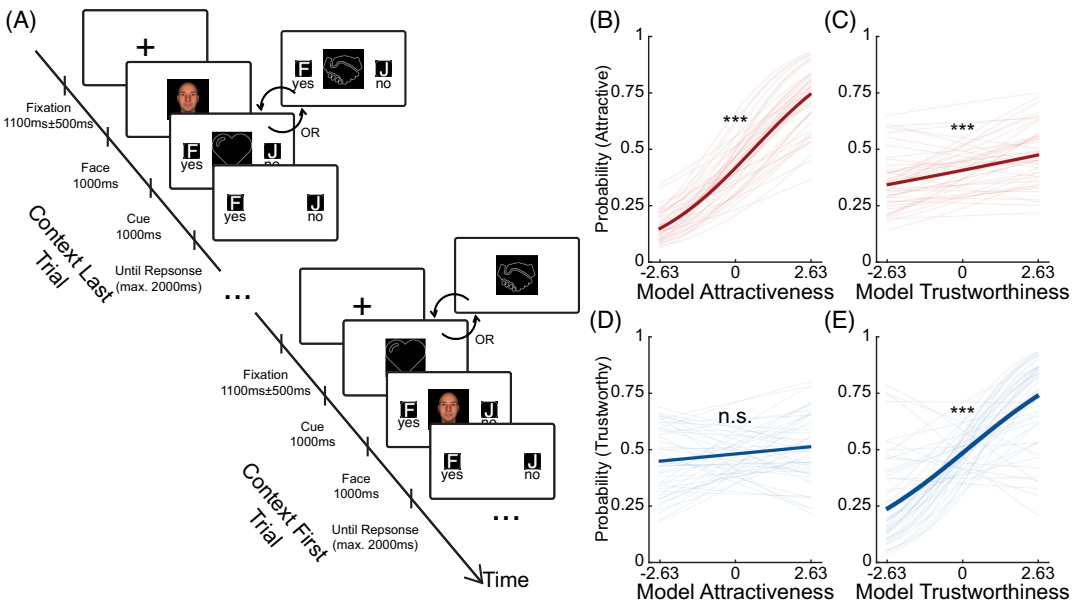
A total of 50 in-lab participants from Brown University community (31 participants reported their genders as female, and 19 as male; average age = 22.4 ± 5.25) participated in Experiment 3. Participants were compensated for their time with a monetary reward. Three participants were removed from the analyses due to their high timeout rates, yielding a final sample size of 47 participants. The inclusion of these three participants does not change our findings.

Design

We used the same design as described in Experiment 2 with one critical difference. Participants experienced two types of trials—context-first and context-last trials—where they were informed about their goal either before or after the face was presented, respectively. Context-first trials adopted an input gating paradigm (essentially, Experiment 2), while context-last trials adopted an output gating paradigm, such that the display order of the visual cue and the face was reversed (Figure 5A). Participants were first presented with a face without knowing which dimension they would be asked about. A heart or handshake icon was then briefly presented, instructing whether participants should predict the judgment on perceived attractiveness or perceived trustworthiness of the face they just saw. Each face ($N = 90$, 15 sets of six faces), derived from the validated model in Experiment 1 was presented twice, once to make predictions on its perceived attractiveness and once for perceived trustworthiness. In total, each participant made 180 predictions, half of which were context-first and half of which were context-last trials; all trials were randomly presented. As in Experiment 2, participants completed a trust game with

Figure 5

Experiment 3: Working Memory Gating Paradigm Design and Results for Context-First Trials



Note. (A) Experimental procedure for context-first and context-last trials in Experiment 3. During context-first trials, participants were given a clear goal of either judging attractiveness or trustworthiness before a face was presented. In context-last trials, participants were presented with a face without a clear goal. After the face disappeared, they were asked to judge its attractiveness or trustworthiness. (B–E) Participants' judgments on social dimensions predicted by relevant and irrelevant information in the context-first trials. The *x*-axis reflects the model estimates of each social attribute (attractiveness and trustworthiness). The *y*-axis denotes the average probability of judging the faces as above average attractiveness (red) or above average trustworthiness (blue). n.s. = not significant. See the online article for the color version of this figure.
*** $p < .001$.

the remaining, unpresented faces ($N = 60$, the other 10 sets of six faces).

Results

Data from context-first trials replicated our findings from Experiment 2: As before, participants judged faces with higher model estimates of attractiveness as more attractive ($\beta_{judgeA\_A} = 0.538 \pm 0.049$, $t = 10.985$, $p < 10^{-26}$; Figure 5B), and faces with higher model estimates of trustworthiness as more trustworthy ($\beta_{judgeT\_T} = 0.419 \pm 0.072$, $t = 5.797$, $p < 10^{-8}$; Figure 5C). Similarly, participants again failed to completely filter out trustworthy information, which ultimately influenced their judgments of other's attractiveness ($\beta_{judgeA\_T} = 0.105 \pm 0.040$, $t = 2.630$, $p = .009$; Figure 5E)—again suggesting a trustworthiness halo effect. Such a halo effect persists even when positive emotion expressions are controlled for ($\beta_{judgeA\_T} = 0.194 \pm 0.050$, $t = 3.856$, $p < 10^{-3}$; Supplemental Figure S4D). As before, participants were able to effectively filter out attractive information, which enabled their judgments of trustworthiness to remain unaffected by any other information ($\beta_{judgeT\_A} = 0.049 \pm 0.046$, $t = 1.072$, $p = .284$; Figure 5D).

Given the natural mapping between context-first trials and the known goal instructed in the trust game, we also anticipated replicating the individual-level modulation effects we observed in Experiment 2. Indeed, participants only relied on trustworthiness information to inform their decisions to trust ($\beta_{modelT} = 0.492 \pm 0.074$, $t = 6.665$, $p < 10^{-10}$; Figure 6A blue line), while not relying on attractiveness information ($\beta_{modelA} = -0.017 \pm 0.056$, $t = -0.299$, $p = .765$; Figure 6A red line). This effect was again modulated by their ability to extract decomposed trustworthiness information ($\beta_{modelT \times \beta_{judgeT\_T}} = 0.808 \pm 0.160$, $t = 5.039$, $p < 10^{-4}$; Figure 6B), which did not extend to attractiveness information ($\beta_{modelT \times \beta_{judgeA\_A}} = 0.250 \pm 0.296$, $t = 0.842$, $p = .405$; Figure 6B).

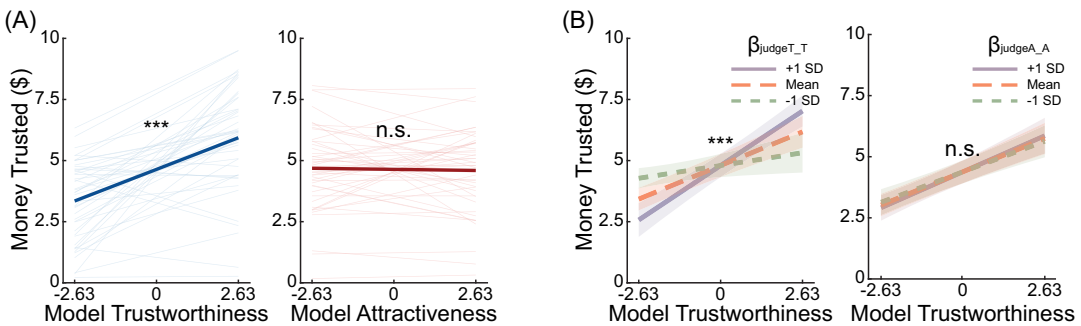
Finally, we used mixed-effects logistic regressions to test participants' predictions about the attractiveness and trustworthiness for both task-relevant and task-irrelevant information in context-last trials. Results revealed that participants were incapable of maintaining factorized representations of attractiveness or trustworthiness, when there was no known goal. Essentially, participants encoded both social dimensions, which affected their judgments of both trustworthiness and attractiveness: Participants judged faces with higher model estimates of attractiveness as more attractive ($\beta_{judgeA\_A} = 0.282 \pm 0.043$, $t = 6.566$, $p < 10^{-10}$; Figure 7B), and they also judged more trustworthy faces as more attractive ($\beta_{judgeA\_T} = 0.249 \pm 0.047$, $t = 5.296$, $p < 10^{-6}$; Figure 7C). Participants' predictions about another's trustworthiness were similarly affected by both irrelevant information (attractiveness: $\beta_{judgeT\_A} = 0.239 \pm 0.038$, $t = 6.237$, $p < 10^{-9}$; Figure 7D) and task-relevant information (trustworthiness: $\beta_{judgeT\_T} = 0.233 \pm 0.048$, $t = 4.809$, $p < 10^{-5}$; Figure 7E)—illustrating an entwined representation. When making trustworthiness judgments, those who exhibited less interference from attractiveness information (Figure 7D, participants with flatter slopes) were able to more effectively use trustworthiness information to inform how much money should be entrusted with other players in the trust game ($\beta_{modelT \times \beta_{judgeT\_A}} = -1.481 \pm 0.601$, $t = -2.463$, $p = .019$; linear regression that includes two interaction terms for model estimates of attractiveness/trustworthiness and the individual-level beta values extracted from the output gating paradigm assessing social attribute judgments).

General Discussion

An open question in face perception is how goals can bias the encoding and representation of social information. By constructing a statistical face model where perceived attractiveness and trustworthiness are truly orthogonal, we created a face set that could be flexibly encoded and represented depending on the context. Mirroring real-life settings where we sometimes have a clear goal, while at other times the environment must be navigated without a goal in mind, we found that people can successfully prevent irrelevant information about a person's attractiveness from affecting their judgments of trustworthiness. Moreover, an individual's ability to factorize across these social dimensions predicted how much money was entrusted to others in a subsequent trust game. That is, deviating from prior work (Jones et al., 2021; Oh et al., 2023; Oosterhof & Todorov, 2008; Stolier et al., 2018), we observed that

Figure 6

Experiment 3: Trust Game Results

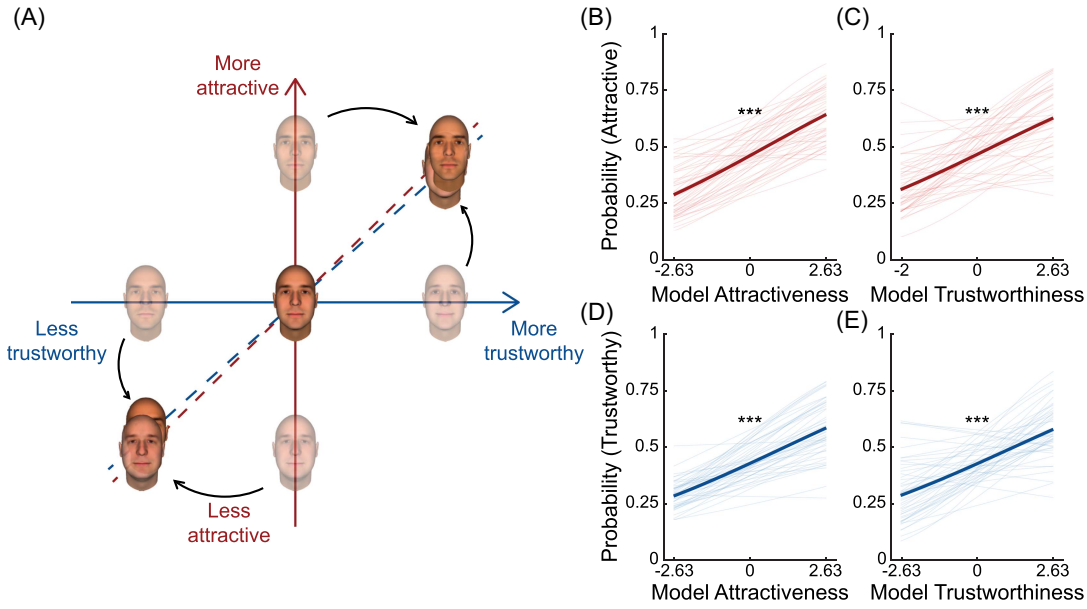


Note. (A) Money entrusted is predicted by model estimated trustworthiness (blue) but not attractiveness (red). (B) This relationship between money entrusted and model estimated trustworthiness is modulated by the degree to which participants were able to rely on the relevant trustworthiness information. n.s. = not significant. See the online article for the color version of this figure.

*** $p < .001$.

Figure 7

Experiment 3: Working Memory Gating Paradigm for Context-Last Trials



Note. (A) A toy illustration of how social dimensions become distorted when both attractive and trustworthy information is encoded in a holistic manner. More attractive faces are perceived to be more trustworthy, and more trustworthy faces are perceived to be more attractive. (B–E) Participants' judgments are predicted by relevant and irrelevant information. The *x*-axis illustrates the model estimates of social attributes (attractiveness and trustworthiness) of the faces. The *y*-axis denotes the average probability of participants judging the faces as above average attractiveness (red) or above average trustworthiness (blue). See the online article for the color version of this figure.

*** $p < .001$.

multidimensional social information can—at times—be successfully decomposed and factorized, which has downstream effects on choice. However, there are also times when information cannot be successfully decomposed and factorized, such as when people are unable to prevent irrelevant information about trustworthiness from influencing their judgments about attractiveness. To better understand this observed trustworthiness bias, we leveraged an output gating paradigm where one does not know beforehand what information might be important since there is no specific goal to meet. In these goal-agnostic environments, people do not seem capable of factorizing multidimensional social information. Instead, they create a gestalt representation that blends perceived attractiveness and trustworthiness, and this holistic representation biases subsequent social judgments.

While our findings illustrate that people are capable of extracting only goal-relevant information when there is a goal in hand, the exact perceptual or cognitive computations that afford such a filtering process remain unknown. One possibility is that the perceptual system adopts an attentional filter to only allow facial features relevant to the goal to pass through and be further processed (Broadbent, 1957; Treisman, 1969), blocking out all other irrelevant information. If this is the case, the observed trustworthy halo effect might reflect a failure in shutting down the "trustworthiness filter" due to its privileged significance to social life (Balliet & Van Lange, 2013; Kramer, 1999; Rotter, 1971, 1980; Schilke et al., 2021). Future research can delineate how specific filters used for different social attributes are acquired and deployed by perceptual systems. In a similar vein, more clarity is needed for characterizing how deeply readouts of social attributes are intertwined in goal-agnostic environment. One possibility is that the perceptual system adopts multiple attentional filters, so that each social attribute is encoded independently early on during perception, but once at the representation stage, these attributes become blended. In addition, the degree to which the representation of one attribute leaks onto another is likely affected by prior beliefs about how these attributes relate to each other (Stolier et al., 2018). Such beliefs are shaped by the natural statistics of personality structures in one's surrounding environment (Oh et al., 2022). Alternatively, it is possible that the perceptual system never applies any attentional filter, when there is no clear goal. Future work can help explore these possibilities underlying the blended representation of social attributes.

The observation that humans can factorize multidimensional social information suggests that people are attempting to simplify the cognitive challenges associated with a high-dimensional problem space, essentially doing a form of dimensionality reduction. Factorization allows for the segmentation of highly complex data into simpler, more manageable components. By extracting the essential features, factorization reduces the overall complexity of the problem to facilitate adaptive learning. For example, if you find yourself rejected after a job interview, a factorized representation of the interview process could help pinpoint the reason(s) behind the failure, in order to aid future improvements (Gershman et al., 2015;

Lamba et al., 2023). In other words, if people can factorize multidimensional information, they can swiftly and efficiently use essential information to behave more adaptively.

To date, understanding whether people can extract only goal-relevant social attributes from faces has been almost impossible given the experimental paradigms the field uses. The result is a literature that assumes that social information is processed in a holistic manner, where distinct social attributes are perceived as closely intertwined (Jones et al., 2021; Oosterhof & Todorov, 2008). Indeed, much research has compiled detailed evidence of how irrelevant social information can bias our judgments, such as how a suspect's attractiveness can lessen the length of a prison sentence (Sigall & Ostrove, 1975; Stewart, 1985). The problem is that once a goal is introduced, there is no way to disentangle the contributions of highly correlated information. One simple solution is to vary the goal state and use a statistical model that orthogonalizes any highly correlated social attributes, such as attractiveness and trustworthiness. Our newly developed face model ensures that social information in one dimension provides zero information along the other dimension. With such methodology in hand, any two social dimensions can be tested against one another to decipher the degree to which different types of information are factorized, or, alternatively, processed in a more gestalt manner. This could include clarifying, for example, whether well-known effects in the literature, such as perceived competence predicting electoral success (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007; Todorov et al., 2005) is due to a direct effect, or instead is because competence is actually biasing perceptions of attractiveness, which indirectly predicts electoral success (a mediating effect; Verhulst et al., 2010).

Our results also demonstrate a strong bias for encoding trustworthiness information; people are unable to filter out trustworthiness information, even when it is irrelevant, which essentially impacts the potency of how other social attributes become represented. That we only observed this along the trustworthiness dimension—but not attractiveness dimension—suggests that trustworthiness should be considered as a privileged social dimension. Such an asymmetrical halo effect might reflect a motivation for prioritizing accurate trustworthiness judgments: Mistrusting an untrustworthy person is far more consequential than misperceiving the physical attractiveness of another person (FeldmanHall et al., 2018). This idea aligns with prior work demonstrating that trustworthiness holds a significant influence in social life (Balliet & Van Lange, 2013; Kramer, 1999; Rotter, 1971, 1980; Schilke et al., 2021) and is often automatically and rapidly processed (Adolphs et al., 1998; Engell et al., 2007; Klapper et al., 2016; Todorov et al., 2008, 2009; Uddenberg et al., 2023; Winston et al., 2002), but it also calls into question other well-known facial processing effects, primarily the long-standing attractiveness "halo effect" (Dion et al., 1972; Eagly et al., 1991; Langlois et al., 2000; van Leeuwen & Neil Macrae, 2004). Given that attractiveness and trustworthiness are typically (if not always) correlated in both natural and artificially generated faces, our results suggest that information about trustworthiness may be the more potent dimension, the one that robustly influences other social attributes (Thorndike, 1920). Indeed, even though we selected what we assumed to be the two most important types of social information, attractiveness and trustworthiness, only attractiveness could be filtered out. Future work can help investigate whether trustworthiness is the only social attribute that is privileged during facial processing, or whether there are other potent social dimensions that function in a similar way.

Depending on the context, people either factorize and filter out irrelevant information or maintain a holistic representation, suggesting a precision–generalization trade-off (Geman et al., 1992). While factorization is premised upon precise encoding of component information, especially when it is relevant to a goal, it also comes at the expense of being able to generalize (FeldmanHall et al., 2018). When there is no clear goal at hand, people seem to maintain a more holistic representation of the available information, since it is not clear which information is relevant for solving future problems. While at first blush such a gestalt representation might seem less advantageous, this format of representation may actually be quite adaptive. Given the vast array of information provided by faces and the limited capacity of human working memory (Cowan, 2012), attempting to factorize and separately represent every piece of information likely requires significant cognitive resources. A holistic or blended representation may be a more resource-efficient strategy that compresses large amounts of information in goal-agnostic environments—situations that are part and parcel of our everyday life (Brady et al., 2009; Franklin & Frank, 2018; Gobet et al., 2001; Nassar et al., 2018; Vives et al., 2023).

In real-world scenarios, the mapping between goals and relevant social cues is rarely straightforward, and instead often involves a complex array of information that relates to a hierarchical objective (FeldmanHall & Nassar, 2021; FeldmanHall & Shenhav, 2019). This complexity may necessitate a gestalt approach to information representation. Consider the example we began with, the foreign intersection where we ask for directions. We are effectively looking for someone who not only appears warm and trustworthy but also seems like a local with knowledge about the nearby environment. If we were to factorize and separately represent all the social information from each stranger's face, we would need to individually extract each dimension for every face in view, according to each singular goal, and only then perform online computations to integrate these dimensions to decide whether to approach a specific stranger. Such a process would be significantly computationally costly. In contrast, a gestalt representation enables semireliable decisions with high efficacy. Therefore, a gestalt representation of social attributes might indeed reflect the adaptive nature of social cognition (Hackel et al., 2024).

These studies took an initial step toward uncovering a flexible encoding mechanism when faced with multidimensional social information. Although our study reveals that people flexibly adjust how they extract social attributes from faces and form representations to fulfill goals, there are unanswered questions. For example, what are the computations humans use to compress multidimensional social information into holistic representation? How do subgoals in a hierarchical social task influence the balance between factorization and gestalt encoding to form a representation with the largest utility that can serve the ultimate goal? Future work can help deepen our understanding of how, and when, factorization might be leveraged in the service of adaptive social behavior.

Constraints on Generality

The first experiment in this set of studies interrogated social evaluations from an online work platform (Amazon Mechnical Turk) using a computer-generated face stimulus set. We constrained this

online sample to fluent English speakers in the United States. Although social evaluations about faces generalize across cultures and regions (Cunningham et al., 1995; Jones et al., 2021), a direct test on a more inclusive population would be necessary to validate how well our statistical face model captures face evaluations writ large. Experiments 2 and 3 took place within the laboratory, and thus we mostly recruited from the general Brown University community. Considering that the variability in personality structure of people living in a region is predictive of the variability in the structure of face impressions (Oh et al., 2022), it remains an open question whether our findings generalize beyond Western cultures. Finally, in order to maintain rigorous experimental control, our findings are based on artificial faces generated with FaceGen. Despite reasonable ecological validity of these faces and their widespread use in face perception studies over decades, future research would benefit from using more naturalistic faces (e.g., hyperrealistic faces generated with generative adversarial networks) to investigate the open questions mentioned above.

References

Abbas, Z.-A., & Duchaine, B. (2008). The role of holistic processing in judgments of facial attractiveness. *Perception*, *37*(8), 1187–1196. https://doi.org/10.1068/p5984

Adolphs, R., Tranel, D., & Damasio, A. R. (1998). The human amygdala in social judgment. *Nature*, *393*(6684), 470–474. https://doi.org/10.1038/30982

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science*, *323*(5918), Article 1183. https://doi.org/10.1126/science.1167748

Asch, S. E. (1946). Forming impressions of personality. *Journal of Abnormal Psychology*, *41*(3), 258–290. https://doi.org/10.1037/h0055756

Ballew, C. C., II, & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(46), 17948–17953. https://doi.org/10.1073/pnas.0705435104

Balliet, D., & Van Lange, P. A. M. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin*, *139*(5), 1090–1112. https://doi.org/10.1037/a0030939

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*(2), 269–278. https://doi.org/10.1037/1528-3542.6.2.269

Baudouin, J.-Y., & Humphreys, G. W. (2006). Configural information in gender categorisation. *Perception*, *35*(4), 531–540. https://doi.org/10.1068/p3403

Bengio, Y., Courville, A., & Vincent, P. (2014). *Representation learning: A review and new perspectives*. PsyArXiv. https://doi.org/10.48550/arXiv.1206.5538

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Bjornsdottir, R. T., Hensel, L. B., Zhan, J., Garrod, O. G. B., Schyns, P. G., & Jack, R. E. (2024). Social class perception is driven by stereotype-related facial features. *Journal of Experimental Psychology: General*, *153*(3), 742–753. https://doi.org/10.1037/xge0001519

Bjornsdottir, R. T., & Rule, N. O. (2020). Negative emotion and perceived social class. *Emotion*, *20*(6), 1031–1041. https://doi.org/10.1037/emo0000613

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* (pp. 187–194). https://doi.org/10.1145/311535.311556

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: Using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487–502. https://doi.org/10.1037/a0016797

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, *64*(3), 205–215. https://doi.org/10.1037/h0047313

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 527–551. https://doi.org/10.1037/0096-1523.26.2.527

Chatham, C. H., Frank, M. J., & Badre, D. (2014). Corticostriatal output gating during selection from working memory. *Neuron*, *81*(4), 930–942. https://doi.org/10.1016/j.neuron.2014.01.002

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., & Abbeel, P. (2016). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems, 29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Abstract.html

Cloutier, J., Mason, M. F., & Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology*, *88*(6), 885–894. https://doi.org/10.1037/0022-3514.88.6.885

Cowan, N. (2012). *Working memory capacity*. Psychology Press. https://doi.org/10.4324/9780203342398

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C.-H. (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, *68*(2), 261–279. https://doi.org/10.1037/0022-3514.68.2.261

Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and Social Psychology*, *24*(3), 285–290. https://doi.org/10.1037/h0033731

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but…: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin*, *110*(1), 109–128. https://doi.org/10.1037/0033-2909.110.1.109

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, *19*(9), 1508–1519. https://doi.org/10.1162/jocn.2007.19.9.1508

Farah, M. J., Wilson, K. D., Drain, M., & Tanaka, J. N. (1998). What is "special" about face perception? *Psychological Review*, *105*(3), 482–498. https://doi.org/10.1037/0033-295X.105.3.482

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*(7), E1690–E1697. https://doi.org/10.1073/pnas.1715227115

FeldmanHall, O., & Nassar, M. R. (2021). The computational challenge of social learning. *Trends in Cognitive Sciences*, *25*(12), 1045–1057. https://doi.org/10.1016/j.tics.2021.09.002

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behaviour*, *3*(5), 426–435. https://doi.org/10.1038/s41562-019-0590-x

Franklin, N. T., & Frank, M. J. (2018). Compositional clustering in task structure learning. *PLOS Computational Biology*, *14*(4), Article e1006116. https://doi.org/10.1371/journal.pcbi.1006116

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

Gershman, S. J., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43–50. https://doi.org/10.1016/j.cobeha.2015.07.007

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in*

Cognitive Sciences, *5*(6), 236–243. https://doi.org/10.1016/S1364-6613(00)01662-4

Hackel, L. M., Kalkstein, D. A., & Mende-Siedlecki, P. (2024). Simplifying social learning. *Trends in Cognitive Sciences*, *28*(5), 428–440. https://doi.org/10.1016/j.tics.2024.01.004

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, *28*(11), 5464–5478. https://doi.org/10.1109/TIP.2019.2916751

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). *beta-VAE: Learning basic visual concepts with a constrained variational framework* [Poster presentation]. 5th International Conference on Learning Representations, ICLR 2017, Toulon, France. https://openreview.net/forum?id=Sy2fzU9gl

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4–5), 411–430. https://doi.org/10.1016/S0893-6080(00)00026-5

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology*, *68*(1), 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Jaeger, B., & Jones, A. L. (2022). Which facial features are central in impression formation? *Social Psychological & Personality Science*, *13*(2), 553–561. https://doi.org/10.1177/19485506211034979

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*(1), 159–169. https://doi.org/10.1038/s41562-020-01007-2

Karras, T., Laine, S., & Aila, T. (2021). A style-based generator architecture for generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(12), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

Kingma, D. P., & Welling, M. (2022). *Auto-encoding variational Bayes*. PsyArXiv. https://doi.org/10.48550/arXiv.1312.6114

Klapper, A., Dotsch, R., van Rooij, I., & Wigboldus, D. H. J. (2016). Do we spontaneously form stable trustworthiness impressions from facial appearance? *Journal of Personality and Social Psychology*, *111*(5), 655–664. https://doi.org/10.1037/pspa0000062

Koffka, K. (1922). Perception: An introduction to the Gestalt-Theorie. *Psychological Bulletin*, *19*(10), 531–585. https://doi.org/10.1037/h0072422

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*(1), 569–598. https://doi.org/10.1146/annurev.psych.50.1.569

Kulkarni, T. D., Whitney, W., Kohli, P., & Tenenbaum, J. B. (2015). *Deep convolutional inverse graphics network*. PsyArXiv. https://doi.org/10.48550/arXiv.1503.03167

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253. https://doi.org/10.1017/S0140525X16001837

Lamba, A., Nassar, M. R., & FeldmanHall, O. (2023). Prefrontal cortex state representations shape human credit assignment. *eLife*, *12*, Article e84888. https://doi.org/10.7554/eLife.84888

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423. https://doi.org/10.1037/0033-2909.126.3.390

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). *Deep learning face attributes in the wild* [Conference session]. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile. https://doi.org/10.1109/ICCV.2015.425

Maestripieri, D., Henry, A., & Nickels, N. (2017). Explaining financial and prosocial biases in favor of attractive people: Interdisciplinary perspectives from economics, social psychology, and evolutionary psychology. *Behavioral and Brain Sciences*, *40*, Article e19. https://doi.org/10.1017/S0140525X16000340

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4

Michel, C., Corneille, O., & Rossion, B. (2007). Race categorization modulates holistic face encoding. *Cognitive Science*, *31*(5), 911–924. https://doi.org/10.1080/03640210701530805

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. *Psychological Review*, *125*(4), 486–511. https://doi.org/10.1037/rev0000101

Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, *30*(1), 65–79. https://doi.org/10.1177/0956797618813092

Oh, D., Martin, J. D., & Freeman, J. B. (2022). Personality across world regions predicts variability in the structure of face impressions. *Psychological Science*, *33*(8), 1240–1256. https://doi.org/10.1177/09567976211072814

Oh, D., Wedel, N., Labbree, B., & Todorov, A. (2023). Trustworthiness judgments without the halo effect: A data-driven computational modeling approach. *Perception*, *52*(8), 590–607. https://doi.org/10.1177/03010066231178489

Olivola, C. Y., & Todorov, A. (2010). Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of Nonverbal Behavior*, *34*(2), 83–110. https://doi.org/10.1007/s10919-009-0082-1

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(32), 11087–11092. https://doi.org/10.1073/pnas.0805664105

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In X. Xie, M. W. Jones, & G. K. L. Tam (Eds.), *Proceedings of the British Machine Vision Conference 2015, BMVC 2015* (pp. 41.1–41.12). British Machine Vision Association Press. https://www.robots.ox.ac.uk/~vgg/publications/2015/Parkhi15/parkhi15.pdf

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 of America*, *119*(17), Article e2115228119. https://doi.org/10.1073/pnas.2115228119

Quinn, K. A., & Macrae, C. N. (2005). Categorizing others: The dynamics of person construal. *Journal of Personality and Social Psychology*, *88*(3), 467–479. https://doi.org/10.1037/0022-3514.88.3.467

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, *26*(5), 443–452. https://doi.org/10.1037/h0031464

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, *35*(1), 1–7. https://doi.org/10.1037/0003-066X.35.1.1

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*(3), 409–426. https://doi.org/10.1037/a0031050

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annual Review of Sociology*, *47*(1), 239–259. https://doi.org/10.1146/annurev-soc-082120-082850

Sigall, H., & Ostrove, N. (1975). Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment. *Journal of Personality and Social Psychology*, *31*(3), 410–414. https://doi.org/10.1037/h0076472

Soulos, P., & Isik, L. (2024). Disentangled deep generative models reveal coding principles of the human face processing network. *PLOS Computational Biology*, *20*(2), Article e1011887. https://doi.org/10.1371/journal.pcbi.1011887

Stewart, J. E., II. (1985). Appearance and punishment: The attraction-leniency effect in the courtroom. *The Journal of Social Psychology*, *125*(3), 373–378. https://doi.org/10.1080/00224545.1985.9922900

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 of the United States of America*, *115*(37), 9210–9215. https://doi.org/10.1073/pnas.1807222115

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*(1), 208–224. https://doi.org/10.1196/annals.1440.012

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: A model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127. https://doi.org/10.1093/scan/nsn009

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*(4), 724–738. https://doi.org/10.1037/a0032335

Todorov, A., Loehr, V., & Oosterhof, N. N. (2010). The obligatory nature of holistic processing of faces in social judgments. *Perception*, *39*(4), 514–532. https://doi.org/10.1068/p6501

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*(5728), 1623–1626. https://doi.org/10.1126/science.1110589

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*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813–833. https://doi.org/10.1521/soco.2009.27.6.813

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, *76*(3), 282–299. https://doi.org/10.1037/h0027242

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, *3*(1), 71–86. https://doi.org/10.1162/jocn.1991.3.1.71

Uddenberg, S., Thompson, B. D., Vlasceanu, M., Griffiths, T. L., & Todorov, A. (2023). Iterated learning reveals stereotypes of facial trustworthiness that propagate in the absence of evidence. *Cognition*, *237*, Article 105452. https://doi.org/10.1016/j.cognition.2023.105452

van 't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition*, *108*(3), 796–803. https://doi.org/10.1016/j.cognition.2008.07.002

van Leeuwen, M. L., & Neil Macrae, C. (2004). Is beautiful always good? Implicit benefits of facial attractiveness. *Social Cognition*, *22*(6), 637–649. https://doi.org/10.1521/soco.22.6.637.54819

Verhulst, B., Lodge, M., & Lavine, H. (2010). The attractiveness halo: Why some candidates are perceived more favorably than others. *Journal of Nonverbal Behavior*, *34*(2), 111–117. https://doi.org/10.1007/s10919-009-0084-z

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 of the United States of America*, *111*(32), E3353–E3361. https://doi.org/10.1073/pnas.1409860111

Vives, M.-L., Frances, C., & Baus, C. (2023). Automating weighing of faces and voices based on cue saliency in trustworthiness impressions. *Scientific Reports*, *13*(1), Article 20037. https://doi.org/10.1038/s41598-023-45471-y

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. *Psychological Bulletin*, *138*(6), 1172–1217. https://doi.org/10.1037/a0029333

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598. https://doi.org/10.1111/j.1467-9280.2006.01750.x

Winston, J. S., Strange, B. A., O'Doherty, J., & Dolan, R. J. (2002). Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, *5*(3), 277–283. https://doi.org/10.1038/nn816

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *2*(1), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception*, *16*(6), 747–759. https://doi.org/10.1068/p160747

Received March 20, 2024
Revision received November 5, 2024
Accepted November 14, 2024 ■