Reverse Correlating Social Face Perception

Ron Dotsch and Alexander Todorov

Social Psychological and Personality Science 2012 3: 562 originally published online 12 December 2011
DOI: 10.1177/1948550611430272

Social Psychological and Personality Science
3(5) 562-571

DOI: 10.1177/1948550611430272
http://spps.sagepub.com

Reverse Correlating Social Face Perception

Ron Dotsch[1] and Alexander Todorov[1]

Abstract

Reverse correlation (RC) techniques provide a data-driven approach to model internal representations in an unconstrained way. Here, we used this approach to model social perception of faces. In the RC task, participants repeatedly selected from two face images—created by superimposing randomly generated noise masks on the same face—the face that looked most trustworthy (or, in other conditions: untrustworthy, dominant, or submissive). We calculated classification images (CIs) by averaging all selected images. Trait judgments of independent participants, as well as objective metrics, showed that the CIs visualized the intended traits well. Furthermore, tests of pixel clusters showed that diagnostic information resided mostly in mouth, eye, eyebrow, and hair regions. The current work shows that RC provides an excellent tool to extract psychologically meaningful images that map onto social perception.

Keywords

facial expressions, measurement, person perception, social cognition, social judgment

It is easy for people to perceive and extract social information from faces. People infer social traits from faces after minimal time exposure (Bar, Neta, & Linz, 2006; Todorov, Pakrashi, & Oosterhof, 2009; Willis & Todorov, 2006). On the other hand, it is hard for people to verbalize what kind of information (i.e., facial feature configurations) they use to make social judgments. There are many reasons for this, including that some diagnostic features may not have verbal labels or that people may not be aware of the cues they use. The problem of finding the facial information that people use in social judgments is further compounded by the fact that the space of possible hypotheses—what features drive specific social perceptions—is infinitely large (Todorov, Dotsch, Wigboldus, & Said, 2011). For example, 15 binary features result in 32,768 combinations, and 20 binary features result in 1,048,576 combinations. Here we show that a data-driven reverse correlation (RC) approach, designed to overcome the aforementioned problems, can be successfully applied to modeling social perception of faces.

In the RC approach, features of stimuli are randomly varied, after which researchers identify which random variations predict judgments. The RC approach originated in the domain of auditory perception during the 70s (Ahumada & Lovell, 1971) and was later adapted for research on vision (Ahumada, 1996, 2002; Beard & Ahumada, 1998; Solomon, 2002) and neurophysiology (Ringach & Shapley, 2004; Victor, 2005). Only recently have researchers begun to apply RC techniques to the study of social perception.

Oosterhof and Todorov (2008) used face space–based RC to identify those features that drive social perception. In their approach, a previously constructed face space, able to represent faces as unique points in a multidimensional space, was used to randomly generate faces, which were rated on two primary social dimensions: trustworthiness and dominance (Todorov, Said, Engell, & Oosterhof, 2008). Based on these ratings, Oosterhof and Todorov constructed models of facial trustworthiness and dominance, consisting of linear combinations of face space dimensions. Moving any face in the direction of, for example, the trustworthiness dimension makes that face look more trustworthy.

The work by Oosterhof and Todorov (2008) demonstrates the applicability of RC methods to social perception. There are some major advantages of using a face space–based RC approach (see Todorov et al., 2011), such as the limited set of variables needed to represent a face, thereby minimizing potential combinations when modeling social dimensions. However, there are some disadvantages. First, the specific face space model used may have affected the Oosterhof and Todorov results. Specifically, the face space model was based on principal component analysis of 271 3D laser scanned faces (FaceGen, Singular Inversions, Toronto, Canada), limiting the representation of faces to linear combinations of the original

[1] Department of Psychology, Princeton University, NJ, USA

Corresponding Author:
Ron Dotsch, Department of Psychology, Princeton University, Princeton, NJ 08540, USA
Email: rdotsch@princeton.edu

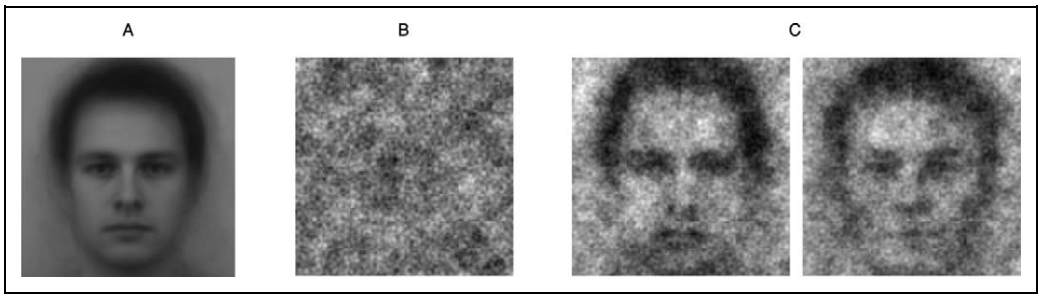

Figure 1. Base face (A), random noise example (B), and example stimuli of noise superimposed on a single base image (C). The left stimulus shows the base image with original noise superimposed and the right stimulus shows the base image with the negative noise superimposed.

faces. Sampling error could have constrained face representation in ways that affect the final models. Second, stimulus faces generated using the face space did not have hair, an important feature affecting person perception (e.g., Macrae & Martin, 2007). Third, only shape information was used to construct the models, although reflectance information (pigmentation and texture) might play an important role in social perception (Todorov & Oosterhof, 2011). Last, although Oosterhof and Todorov included visualizations of the constructed social dimensions, they did not report quantitative analyses with respect to diagnostic face regions.

Here, we use a different class of RC techniques aimed at extending the findings of Oosterhof and Todorov (2008) by assessing whether their findings are method invariant and by quantitatively identifying which specific facial regions are involved in social perception. This class of RC techniques was developed in parallel by Kontsevich and Tyler (2004) and Mangini and Biederman (2004, also see Gosselin & Schyns, 2003). This technique enables researchers to generate images that reflect participants' internal representations of faces, without making any assumption about what those representations might look like. Recently, this RC variant has become increasingly popular in social cognitive research (see, e.g., Dotsch, Wigboldus, Langner, & van Knippenberg, 2008; Dotsch, Wigboldus, & van Knippenberg, 2011; Imhoff, Dotsch, Bianchi, Banse, & Wigboldus, in press; Jack, Caldara, & Schyns, 2011; Karremans, Dotsch, & Corneille, in press).

A typical RC image classification task (in this example we discuss a two images forced choice, or 2IFC, variant used by Dotsch et al., 2008) employs random variations of facial images created with a constant base face (Figure 1A) and randomly generated noise patterns (Figure 1B) superimposed on the face. Because the noise distorts the base face image, the face looks different with each different random noise pattern. For each superimposed random noise pattern, a negative pattern (the mathematical opposite) is generated. Each pixel that is dark in the original noise pattern is bright in the negative noise pattern, much like photo negatives. In a single trial, the base image with the original noise and the base image with the negative noise superimposed are presented side by side (Figure 1C). Participants are then asked to select the face that best resembles the target category. The average of all selected noise patterns constitutes the classification image (CI), whereas the average of all unselected noise patterns is the anti-CI. For instance, Dotsch, Wigboldus, Langner, and van Knippenberg (2008) used a neutral male face as base face, and superimposed random noise consisting of multiple truncated sinusoids. These were then used as stimuli in a Moroccan classification task: participants chose from two stimuli (Figure 1C) the stimulus that best resembled a Moroccan face. The average of all noise patterns that participants classified as Moroccan constituted the Moroccan CI. Superimposing this CI on top of the original base image resulted in approximations of what participants thought typical Moroccan faces looked like.

The 2IFC task described above is based on the psychophysical RC methodology described by Mangini and Biederman (2004), where one image was presented in each trial and participants classified the image into one of two categories. In this case, CIs for each category are calculated by averaging all images classified as the respective category. Mangini and Biederman demonstrated that this technique can be used to model identities (John Travolta vs. Tom Cruise), gender categories (male vs. female), and emotional expressions (happy vs. sad). However, the extent to which this task can be applied to modeling social perception is somewhat limited. First, the task used as base face a morph between two images that accurately represented the two target categories (e.g., John Travolta's and Tom Cruise's face). When modeling social dimensions, researchers do not possess images that accurately represent the two target categories. Instead, those images are exactly what researchers set out to discover. Second, participants discriminated between two specific categories, making it impossible to tap into the internal representation of just one category without contrasting it with another category. The 2IFC variant does not suffer from these problems, because participants select the image that best fits one target category and the images can be created using a base face unrelated to the target category.

Here, we use the 2IFC RC image classification task to model perception of face trustworthiness and dominance. Importantly, the 2IFC task enables us to go beyond the results of Oosterhof and Todorov (2008) in two ways. First, the base image of stimulus faces can include hair and the superimposed noise affects shape as well as reflectance information. Second, with the 2IFC task we can identify facial regions diagnostic for social perception. Arguably, this might be achieved using yet another RC technique, bubbles (Gosselin & Schyns, 2001), in

which good exemplars of the target categories are partially presented (masked by Gaussian "bubbles"). RC analysis then reveals which facial areas predict the correct response. However, the known good exemplars of trustworthiness and dominance have been visualized by Oosterhof and Todorov but might not be optimal because of the reasons mentioned above. The current method, on the other hand, does not need good exemplars as stimuli, but nevertheless can be used to identify diagnostic facial regions.

On the methodological side, the current study provides the opportunity to investigate the interpretation of anti-CIs. Because as yet their meaning is unclear, we included additional RC tasks to validate the anti-CIs. It is likely that an anti-CI will resemble the other end of a bipolar target dimension. For example, in a trustworthiness RC task, the anti-CI may look untrustworthy. We therefore included both trustworthiness and untrustworthiness RC tasks to quantify the extent to which the untrustworthy CI is similar to the antitrustworthy CI and vice versa. For the same reason, we included both dominance and submissiveness RC tasks.

In sum, participants completed a trustworthiness, untrustworthiness, dominance, or submissiveness 2IFC RC task. For example, in the trustworthiness task, on each trial, participants decided which image looked more trustworthy. We calculated the resulting CIs and anti-CIs and used a combination of objective similarity metrics and subjective ratings of independent judges to assess their validity. Because the noise patterns in this task consisted of variations in different regions of the face, we explored which facial regions contained information diagnostic for social judgments using clusters of pixels tests (analogous to tests used with functional magnetic resonance imaging research to identify activated brain regions; see Chauvin, Worsley, Schyns, Arguin, & Gosselin, 2005).

Method

Participants and Design

Eighty students from Princeton University participated. Each performed a single RC task for a trait that was either indicative (e.g., trustworthy and dominant) or counterindicative (e.g., untrustworthy and submissive) for one dimension (respectively, trustworthiness or dominance). We generated two stimulus sets for the RC tasks. Participants were either presented with one or the other stimulus set. Because both sets resulted in highly similar results, we collapsed across this variable. The RC study thus consisted of a 2 (Dimension: trustworthiness vs. dominance) $\times$ 2 (Trait: indicative vs. counterindicative) between-subject design, with 20 participants per cell.

Materials and Procedure

The stimuli in the RC task all consisted of the same base face with different superimposed random noise on each trial. The base face was a gray scale average of all male faces in the Karolinska Face Database (Lundqvist, Flykt, & Öhman, 1998, see Figure 1A). The noise consisted of superimposed truncated sinusoid patches of 2 Cycles in 6 Orientations (0°, 30°, 60°, 90°, 120°, and 150°) $\times$ 5 Spatial scales (2, 4, 8, 16, and 32 cycles per image) $\times$ 2 Phases (0, π/2), with random contrasts (see Figure 2). In sum, the random noise was a function of 4,092 parameters, each defining the contrast value of one truncated sinusoid spanning two cycles. Stimulus size was 512 $\times$ 512 pixels.

In a single trial two stimuli were presented side by side. One stimulus was the base face with a random noise pattern superimposed and the other base face with the negative of the random noise pattern superimposed (see Figure 1C). We chose to use the negative of the random noise pattern as opposed to just another random noise pattern to maximize the differences between the two presented images, to minimize the number of possible stimulus pairs to be presented, and to simplify data analysis. This procedure has been successfully employed by Dotsch et al. (2008), Dotsch, Wigboldus, and van Knippenberg (2011), Imhoff, Dotsch, Bianchi, Banse, and Wigboldus (in press), and Karremans, Dotsch, and Corneille (in press). Participants were instructed to select the stimulus that most resembled a trustworthy (untrustworthy, dominant, and submissive) face.

Participants completed 300 trials. The presented stimuli were drawn without replacement from one of two sets of 300 original (and 300 matching negative) noise patterns. These two sets were generated to ensure that the obtained results were not an artifact of the specific sets of noise patterns used. The stimulus pairs were presented in random order. The placement of the facial images with original and negative noise on the screen (negative noise on the left vs. on the right) was counterbalanced across trials. A 1,000-ms centered fixation cross preceded each trial. After participants completed the task, they were debriefed.

Data Processing

To generate the CIs, we calculated the mean of all noise patterns a participant selected as most trustworthy (untrustworthy, dominant, and submissive), by averaging the parameters on which those noise patterns were based. This resulted in 4,092 mean parameters per participant. We then averaged the mean parameters across participants for each cell of the design and generated the classification patterns based on cell average parameters. Finally, we superimposed the classification patterns on the original base image to generate the CIs. The faces participants *did not* select as most trustworthy (untrustworthy, dominant, and submissive) underwent the same treatment and yielded the anti-CIs.

Results

The resulting CIs and anti-CIs per trait are depicted in Figure 3. Visual inspection of the CIs show that a trustworthy face involves a smooth, small face, a smiling mouth, and open eyes, whereas an untrustworthy face seems to involve a downturned mouth with thick lips, angry-looking eyes, sagging cheeks, and a bald spot on top of the head. The dominant face has clear eyebrows, dark eyes, and a slightly downturned mouth. The face also emerges more from the background. The submissive face

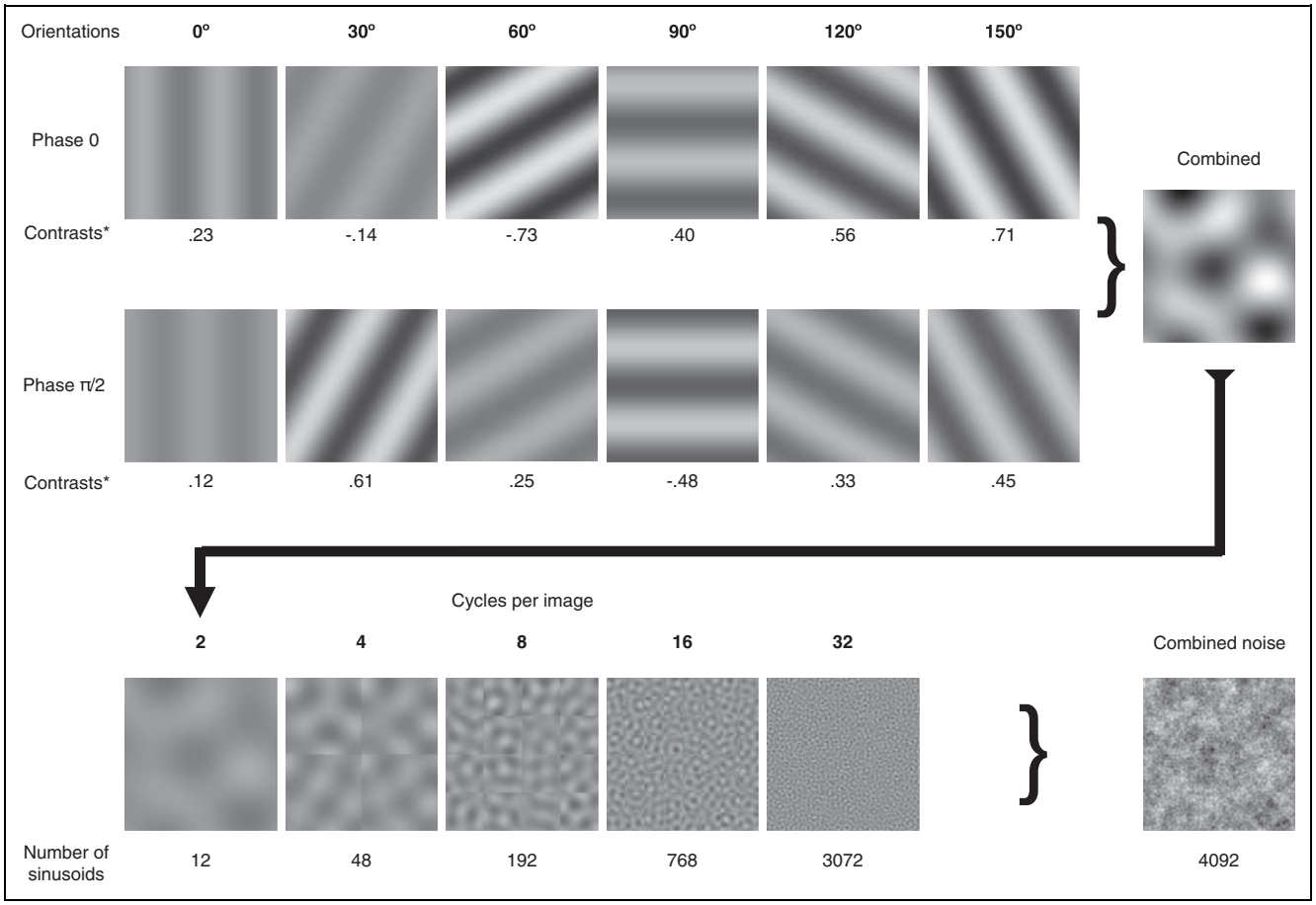

Figure 2. This figure outlines the process of generating the noise pattern that was superimposed on the base image. For each spatial frequency (2, 4, 8, 16, and 32 cycles per image), 12 sinusoids per cycle were superimposed (6 Orientations × 2 Phases). Each sinusoid had its own random parameter indicating its contrast. The first 2 lines indicate how 12 sinusoids form the combined noise for 1 spatial frequency (2 cycles per image). The last line indicates how sinusoids across all spatial frequencies are combined to create the resulting noise. *Note.* *Contrasts are random example values to illustrate how contrast values relate to the resulting sinusoid patches.

is frowning, has thin lips and sad-looking eyes. Moreover, the face delineates itself less from the background.

Participant Agreement

To assess participant agreement, we calculated Cronbach's α on the pixel luminance values of each participants' CI noise pattern (masked by an oval shape to only include pixels of the face), for each social judgment. α, in this analysis, reflects the extent to which the pixels of individual participants' CIs within a cell covary with the pixels of all other participants' CIs in that cell. It was not possible to assess agreement based on the decision data, because two different sets of stimuli were used. As can be seen in Table 1, there is fair agreement with α ranging between .56 and .76.

Objective Metrics

Because the CIs represent psychologically meaningful constructs that have well-defined mutual relations (e.g., trustworthiness is the opposite of untrustworthiness, and relatively unrelated to dominance or submissiveness), the resulting CIs should have the same mutual relations (e.g., the trustworthy CI should be similar to the opposite of the untrustworthy CI, and not too similar to the dominance or submissive CIs). These relations can be quantified by calculating the correlation between the pixel luminance values of one classification pattern and the pixel luminance values of another classification pattern. These correlations can be best understood as measures of similarity. A high-positive correlation means that the CIs are physically similar, a high-negative correlation means that the CIs are physically opposite, and a zero correlation means that the CIs have little in common. These correlations (based on CIs masked with an oval shape to only include pixels of the face) are summarized in Table 2.[1]

The correlations in Table 2 show that (1) CIs for a trait on one side of the dimension (e.g., trustworthy) are physically *different* from (i.e., correlate negatively with) CIs for a trait on the other side of the dimension (e.g., untrustworthy); (2) CIs for a trait on one side of the dimension (e.g., trustworthy) are physically *similar* to (i.e., correlate positively with) CIs of the antiface of a trait on the other side of the dimension (e.g., anti-

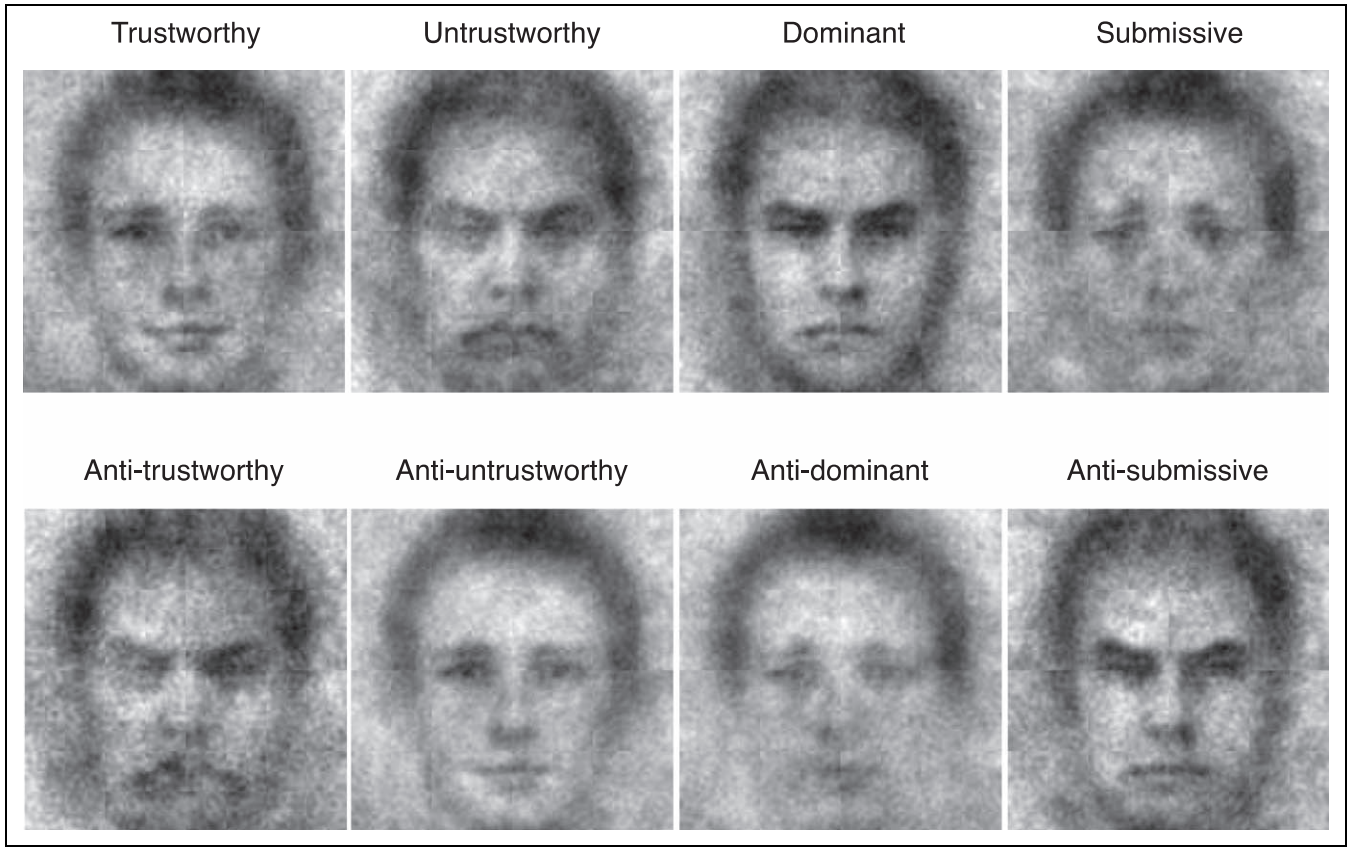

Figure 3. Resulting classification images ([CIs] top row, the average of all noise patterns selected as best resembling the target trait, superimposed on the base image) and anti-CIs (bottom row, the average of all noise patterns *not* selected as best resembling the target trait, superimposed on the base image).

Table 1. Participants' Agreement Measures for Each Social Judgment Quantified as Cronbach's α Computed Over Subjects' CI Noise Pattern Pixel Luminance Values

CI	Cronbach's α
Trustworthy	.56
Untrustworthy	.65
Dominant	.76
Submissive	.58

Table 2. Objective Metric of Similarity Between the Various Visualized Traits Quantified as Correlations Between the Aggregated CI Noise Patterns Pixel Values by Trait

	1	2	3	4
1. Trustworthy				
2. Untrustworthy	−.65			
3. Dominant	−.27	.50		
4. Submissive	.23	−.41	−.70	
5. Antiface[a]	.66	.67	.71	.71

[a]The antiface is the CI based on the unselected faces in the reverse correlation task for the trait on the other end of the same dimension (e.g., the trustworthy face is correlated with the anti-untrustworthy face).

untrustworthy); and (3) CIs for a trait on one dimension (e.g., trustworthiness) are *unrelated* (weakly correlated) to CIs pertaining to another dimension (e.g., dominance). Note however that the correlations with unrelated dimensions are nonzero, indicating that trustworthiness and dominance, although unique dimensions, are not completely orthogonal.

Subjective Metrics

As subjective metric, an independent sample of 38 Princeton University undergraduate students were asked to rate the trustworthy, untrustworthy, dominant, and submissive CIs (one aggregate CI per trait) on trustworthiness, dominance, and threat, on a scale from 1 (*not trustworthy* [*dominant, threatening*]) to 9 (*very trustworthy* [*dominant, threatening*]). Participants rated the stimuli in random order within blocks (one type of judgment per block). The blocks had a fixed order, respectively, trustworthiness, dominance, and threat. We added threat because previous work by Oosterhof and Todorov (2008) showed that threat is a combination of untrustworthiness and dominance. Because both dimensions contribute to threat, we expected the untrustworthy and dominant CIs to be rated equally high on threat, and the trustworthy and submissive CIs to be rated equally low on threat. Beforehand, to prevent that participants would be unfamiliar with the faces when judging

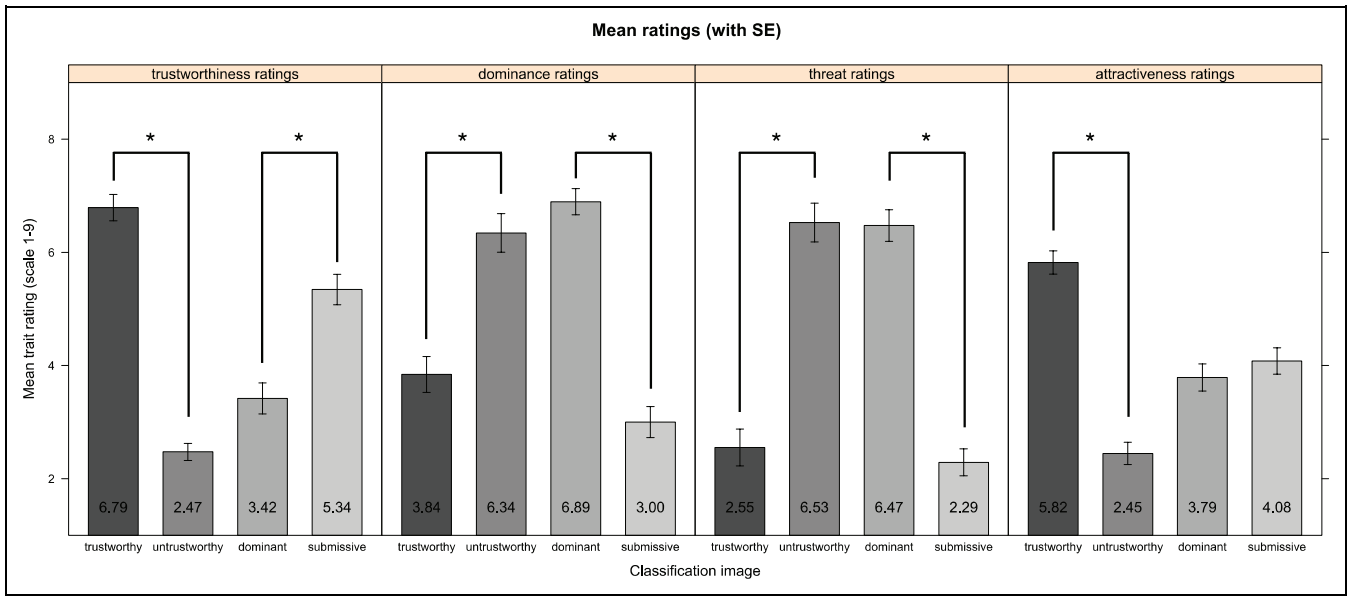

Figure 4. Mean trustworthiness, dominance, threat, and attractiveness ratings of classification images ([CIs] *$p < .05$).

trustworthiness and familiar when judging dominance and threat, participants rated the stimuli on attractiveness to become familiar with the stimulus set.

he average ratings are summarized in Figure 4 and indicate that all CIs visualized the intended traits. That is, the trustworthy CI was judged most trustworthy, whereas the untrustworthy CI was judged least trustworthy. Likewise, the dominant CI was judged most dominant, whereas the submissive CIs were judged least dominant. We subjected the ratings to a 2 (Dimension: trustworthiness vs. submissiveness[2]) × 2 (Trait: indicative vs. counterindicative) within-subject analysis of variance (ANOVA) for each separate trait. The ANOVA on the trustworthiness ratings revealed both a main effect for trait, $F(1, 37) = 122.94$, $p < .001$, $\eta_G^2 = .54$, indicating that the trustworthy and submissive CIs were rated more trustworthy than the untrustworthy and dominant CIs, and a Dimension × Trait interaction effect, $F(1, 37) = 50.06$, $p < .001$, $\eta_G^2 = .15$, indicating that the trustworthiness difference between trustworthy and untrustworthy was greater than the difference between submissive and dominance. The ANOVA on the dominance ratings again revealed both a main effect for trait, $F(1, 37) = 51.29$, $p < .001$, $\eta_G^2 = .44$, indicating that the trustworthy and submissive CIs were rated less dominant than the untrustworthy and dominant CIs, and a Dimension × Trait interaction effect, $F(1, 37) = 11.54$, $p < .001$, $\eta_G^2 = .04$, indicating that the dominance difference between trustworthy and untrustworthy was smaller than the dominance difference between submissive and dominance. Last, as expected, the ANOVA on the threat ratings revealed only a main effect for trait, $F(1, 37) = 67.07$, $p < .001$, $\eta_G^2 = .56$, indicating that the trustworthy and submissive CIs were rated less threatening than the untrustworthy and dominant CIs, and no other differences were significant, n.s.[3]

To test whether judgments of both trustworthiness and dominance contributed to judgments of threat, we compared goodness of fit (R^2) of simple regression models (in which trustworthiness or dominance ratings predicted threat ratings in separate models) with an additive model (in which both trustworthiness and dominance were included as predictors of threat ratings). The additive model predicted threat ratings better ($R^2 = .73$, $\beta_{trustworthiness} = -.35$, $p < .001$; $\beta_{dominance} = .64$, $p < .001$) than trustworthiness ($R^2 = .41$, $\beta_{trustworthiness} = -.64$, $p < .001$) or dominance ratings alone ($R^2 = .64$; $\beta_{dominance} = .80$, $p < .001$), with significant improvements of fit, respectively, $F(1, 149) = 177.75$, $p < .001$, and $F(1, 149) = 52.42$, $p < .001$.

Diagnostic Facial Regions

We performed a cluster test (Chauvin et al., 2005) on the pixel data of each CI's noise pattern to identify facial regions diagnostic for making the various social judgments. We first smoothed the CI noise pattern using a Gaussian filter ($\sigma_b = 4$ pixels). The smoothed image was masked by an oval shape (only revealing pixels located in the face, including hair) and Z transformed. We then performed two-tailed cluster tests (using the stat4CI toolbox, Chauvin et al., 2005; $Z_{crit} \geq |2.3|$, $p < .05$) for each CI to identify clusters containing pixels for which luminance variation predicted social judgments. The resulting clusters, representing facial regions diagnostic for the respective social judgment, are depicted in Figure 5. The clusters in green indicate that pixel luminance in those clusters correlated positively with that respective classification (when these pixels were lighter, participants were more likely to select the image), whereas clusters in red indicate that pixel luminance in those clusters correlated negatively with that respective classification (when these pixels were darker, participants were more likely to select the image).

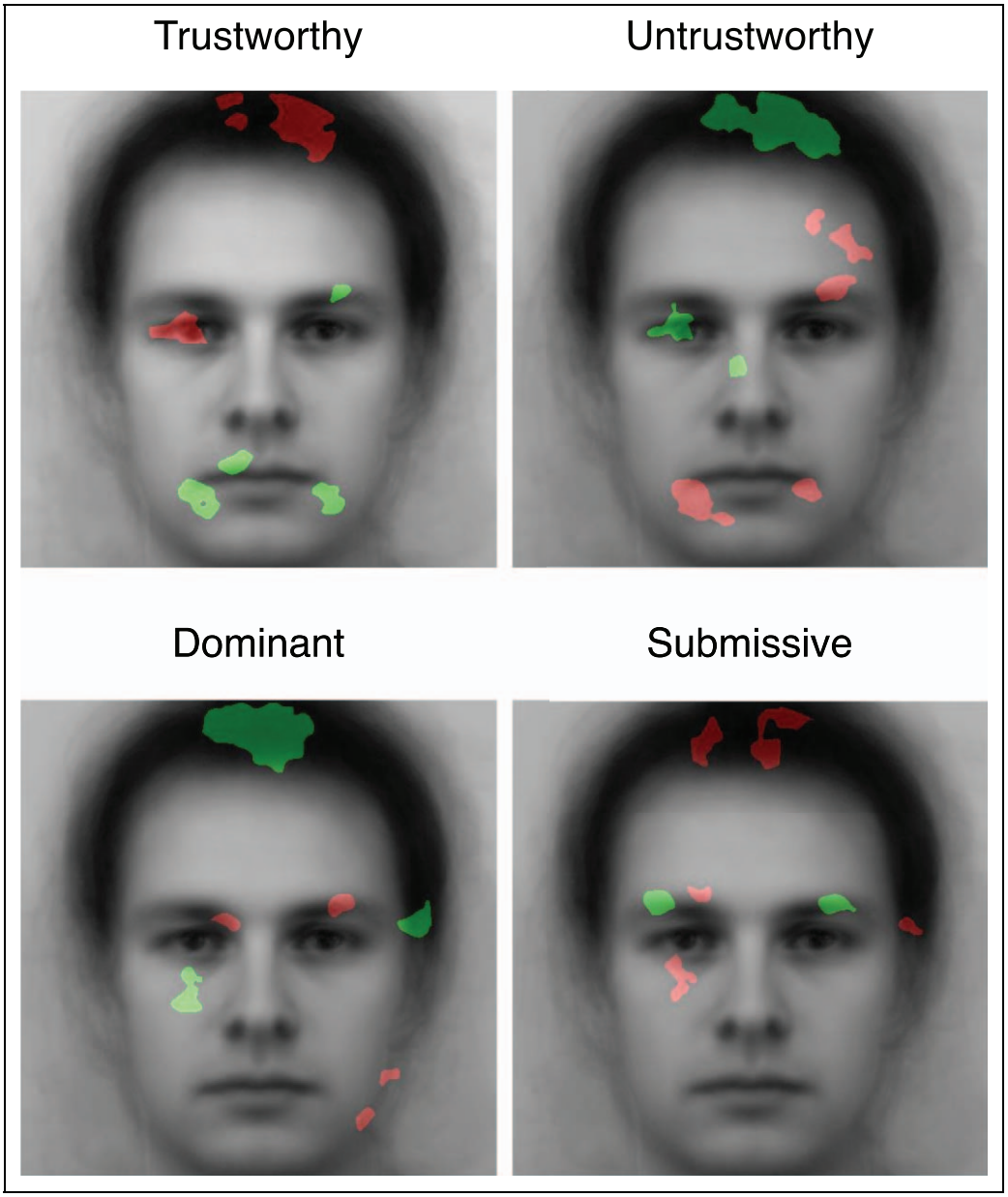

Figure 5. Significant clusters in classification images. Green clusters show where pixel luminance positively predicted classifications, whereas red clusters show where pixels luminance negatively predicted classifications.

The clusters in Figure 5 show that for trustworthiness and untrustworthiness judgments mouth, eye, and hair regions were most important. Whereas trustworthiness included bright areas in the mouth and eyebrow regions and dark areas in the eye and hair regions, the opposite held for untrustworthiness. Likewise, the clusters show that for dominance and submissiveness judgments areas around the eyes, hair regions, and delineation of the face (most visible in the right chin of the dominant CI) were diagnostic.

Discussion

Psychophysical RC provides a data-driven approach to model social perception. Here, we used a 2IFC RC image classification task (Dotsch et al., 2008) to model perception of two primary dimensions of face evaluation: trustworthiness and dominance. To validate the resulting visualizations, we used correlations between pixels of the CIs as objective metrics of similarity. The negative correlation between the resulting CIs for trustworthy (dominant) and untrustworthy (submissive) faces indicated that features that increase face trustworthiness (dominance) decrease face untrustworthiness (submissiveness). The weak positive correlation between the CIs involving the trustworthiness dimension and the CIs involving the dominance dimension established discriminant validity in the sense that features that changed face trustworthiness did not change face dominance much. Moreover,

the high correlations between CIs on one side of the dimension (trustworthy/dominant) and the anti-CIs on the other side of the dimension (untrustworthy/submissive) provided evidence for the interpretation of the anti-CIs as visualizing the negative side of the intended dimension. Furthermore, using subjective ratings of the CIs by independent participants, we showed that the CIs did represent the intended specific social traits. In short, the 2IFC RC image classification task is capable of visualizing social dimensions, other than gender, identity, and emotional expression.

We explored the regions of the face that were diagnostic for the various social judgments (see Figure 5), which highlights the rich body of data that can be gained from RC tasks. Variation in mouth, eye, eyebrow, and hair regions significantly predicted social judgment, which converges with models of trustworthiness and dominance developed by Oosterhof and Todorov (2008), although they used different methods. On the other hand, our analysis goes beyond the findings of Oosterhof and Todorov by identifying diagnostic regions that have been previously ignored. These included hair regions which suggest that hairstyle might be an important cue for making trustworthiness and dominance judgments. Furthermore, the RC analysis revealed that dominant faces might be perceived as more protruding from the background, whereas submissive faces seemed to be more part of the background.

The subjective judgments of the CIs showed a very clear pattern: When judged on trustworthiness, the difference between the trustworthy and untrustworthy CIs was larger than the difference between the dominant and submissive CIs, whereas the opposite was the case when judged on dominance. Importantly, when judged on threat (a trait with both an untrustworthy and a dominant component, see Oosterhof & Todorov, 2008), the difference between trustworthy and untrustworthy CIs was equal to the difference between dominant and submissive CIs. Moreover, judgments of both trustworthiness and dominance contributed to judgment of threat. Although these data provide clear support for the validity of the RC approach to model social perception, two minor artifacts in the data should be addressed. First, the difference between trustworthy and untrustworthy (dominant and submissive) CIs was not absent in dominance (trustworthiness) judgments, indicating that the trustworthiness and dominance dimensions were not completely orthogonal. This is consistent with Oosterhof and Todorov (2008) who also found that the two dimensions were not completely orthogonal. Second, it seems that trustworthiness judgments differentiated better between the two dimensions than dominance judgments. This might be due to fixed ordering; CIs were always rated on trustworthiness before they were rated on dominance, potentially causing participants to cluster some CIs together. Alternatively, both artifacts may be interpreted in light of work on the compensation hypothesis (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005; Yzerbyt, Kervin, & Judd, 2008; Yzerbyt, Provost, & Corneille, 2005) which holds that people perceived as high on one of the two primary social dimensions are perceived as low on the other dimension (also see Fiske, Cuddy, & Glick, 2007). This effect is evident in Figure 4: the trustworthy CI was judged less dominant than the untrustworthy CI, and the dominant CI was judged less trustworthy than the submissive CI. Moreover, the finding that trustworthiness judgments differentiated between dimensions better than dominance is in line with work showing that people are primarily oriented toward warmth or valence information (Wojciszke, 2005) and as a result show stronger compensation effects on the second social dimension (in this case, dominance; Yzerbyt et al., 2008). It should be noted, however, that most work on compensation focused on competence instead of dominance as a second dimension. Whether dominance and competence can be considered different labels for the same social dimension is unclear.

It is important to note that the CIs should not be interpreted as actual mental representations. They are approximations influenced by task-specific factors such as base image or used noise patterns. Nonetheless, the current work clearly shows that RC methods can extract psychologically meaningful images that map onto social perception in a completely unconstrained fashion. The psychological constructs are not necessarily limited to the social dimensions visualized here, nor are they limited to gender, identity, and emotional expressions (Mangini & Biederman, 2004). The possibilities for laying bare internal representations are endless. Jack, Caldara, and Schyns (2011) used the method to identify cultural differences in the perception of emotional expressions. Dotsch et al. (2011) have used the method to demonstrate biases in the representation of social categories. Karremans et al. (in press) used it to uncover memory biases in the representation of attractive potential mates, and Imhoff et al. (in press) to reveal spontaneous ingroup projection in the facial domain. In the future, the method might even prove to have applied value for domains such as eyewitness testimony (as means to create composites of perpetrators, with the possibility of combining data from multiple eyewitnesses). It is our hope that this research paves the way for more psychophysical RC projects to emerge in the social domain.

Acknowledgments

We thank Jenny Porter for her help with the reverse correlation task and Hillel Aviezer for his help with the rating task.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: National Science Foundation grant 0823749, the Russell Sage Foundation, and Netherlands Organisation for Scientific Research (NWO) and Marie Curie Cofund Action Rubicon grant, 446-10-014.

Notes

1. It is also possible to correlate parameters instead of pixels. However, since some parameters (i.e., low spatial frequency parameters) affect many pixels while being few in number, and other parameters (i.e., high spatial frequency parameters) affect just a few pixels in the image while being many in number, the resulting correlations are biased toward high spatial frequency similarity. Moreover, using pixels allowed us to mask those pixels that are not positioned on the face, making the statistics more precise. Nonetheless, we also calculated parameter correlations and the resulting correlation matrix showed a similar pattern.
2. The reason for recoding dominance to submissiveness in these analyses is ease of interpretation, as given in Figure 5: a significant interaction effect now indicates that the difference between trustworthiness and untrustworthiness is larger (or smaller) than the difference between dominant and submissive.
3. Although not pertaining to our current purposes, Figure 4 depicts the attractiveness ratings too. These ratings preceded all other ratings and were included to familiarize participants with the stimuli. We had no prior expectations about the attractiveness ratings, but as can be seen in Figure 4, they closely mimic the trustworthiness ratings. This is in line with the findings of Oosterhof and Todorov (2008) that trustworthiness is highly correlated with several other positive dimensions such as attractiveness and might be described as approximating general valence dimension.

References

Ahumada, A. J. (1996). Perceptual classification images from Vernier acuity masked noise. *Perception*, *26*, 1831–1840.

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*, 121–131.

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustical Society of America*, *49*, 1751–1756.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion*, *6*, 269–278.

Beard, B. L., & Ahumada, A. J. (1998). A technique to extract relevant image features for visual tasks. In *Proceedings of SPIE Human Vision and Electronic Imaging III* (Vol. *3299*, pp. 79–85).

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005). Accurate statistical tests for smooth classification images. *Journal of Vision*, *5*, 1–1. doi: 10.1167/5.9.1

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

Dotsch, R., Wigboldus, D. H. J., & van Knippenberg, A. (2011, March 28). Biased allocation of faces to social categories. *Journal of Personality and Social Psychology*. Advance online publication. doi: 10.1037/a0023026

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*, 77–83.

Gosselin, F., & Schyns, P. G. (2001). Bubbles: A technique to reveal the use of information in recognition tasks. *Vision Research*, *41*, 2261–2271.

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

Imhoff, R., Dotsch, R., Bianchi, M., Banse, R., & Wigboldus, D. H. J. (in press). Facing Europe: Visualizing spontaneous ingroup projection. *Psychological Science*.

Jack, R. E., Caldara, R., & Schyns, P. G. (2011). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General.* doi: 10.1037/a0023463

Judd, C. M., James-Hawkins, L., Yzerbyt, V. Y., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*, 899–913.

Karremans, J. C., Dotsch, R., & Corneille, O. (in press). Romantic relationship status biases mental representations of attractive opposite-sex others: Evidence from a reverse-correlation paradigm. *Cognition*.

Kontsevich, L., & Tyler, C. W. (2004). What makes Mona Lisa smile. *Vision research*, *44*, 1493–1498. doi: 10.1016/j.visres.2003.11.027

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska Directed Emotional Faces.* Stockholm, Sweden: Psychology Section, Department of Clinical Neuroscience, Karolinska Institute.

Lundqvist, D., & Litton, J. E. (1998). The Averaged Karolinska Directed Emotional Faces-AKDEF, CD ROM from department of clinical neuroscience, psychology section, Karolinska Institutet (Tech. Rep.). ISBN 91-630-7164-9.

Macrae, C. N., & Martin, D. (2007). A boy primed sue: Feature-based processing and person construal. *European Journal of Social Psychology*, *37*, 793–805.

Mangini, M. C., & Biederman, I. (2004). Making the ineffable explicit: Estimating the information employed for face classifications. *Cognitive Science*, *28*, 209–226.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the USA*, *105*, 11087–11092.

Ringach, D., & Shapley, R. (2004). RC in neurophysiology. *Cognitive Science*, *28*, 147–166.

Solomon, J. A. (2002). Noise reveals visual mechanisms of detection and discrimination. *Journal of Vision*, *2*, 105–120.

Todorov, A., Dotsch, R., Wigboldus, D. H. J., & Said, C. P. (2011). Data-driven methods for modeling social perception. *Social and Personality Psychology Compass, 5* (10), 775–791. doi: 10.1111/j.1751-9004.2011.00389.x

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces. *Signal Processing Magazine, IEEE, 28*, 117–122.

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

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.

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience*, *8*, 1651–1656.

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

Wojciszke, B. (2005). Morality and competence in person and self perception. *European Review of Social Psychology*, *16*, 155–188.

Yzerbyt, V. Y., Kervyn, N., & Judd, C. M. (2008). Compensation versus halo: The unique relations between the fundamental dimensions of social judgment. *Personality and Social Psychology Bulletin*, *34*, 1110–1123.

Yzerbyt, V. Y., Provost, V., & Corneille, O. (2005). Not competent but warm. Really? Compensatory stereotypes in the French-speaking world. *Group Processes and Intergroup Relations*, *8*, 291–308.

Bios

Ron Dotsch is a postdoctoral researcher at the department of psychology at Princeton University. He is currently working with Alexander Todorov to study social face perception. He received his PhD in Psychology from Radboud University Nijmegen, where he worked with Daniel Wigboldus and Ad van Knippenberg. Ron's research interests include face perception, social categorization, prejudice, stereotypes, and the use of advanced methods, such as virtual reality, reverse correlation, and face space models.

Alexander Todorov is an associate professor of psychology and public affairs at Princeton University with a joint appointment in the Department of Psychology and the Woodrow Wilson School of Public and International Affairs. He is also an affiliated faculty of the Princeton Neuroscience Institute and a visiting professor at Radboud University Nijmegen in the Netherlands. His research focuses on the cognitive and neural basis of social cognition. His main line of research is on the cognitive and neural mechanisms of person perception with a particular emphasis on the social dimensions of face perception. His research approach is multidisciplinary, using a variety of methods from behavioral and functional Magnetic Resonance Imaging experiments to computer and statistical modeling. Alexander received his PhD in psychology from New York University in 2002.